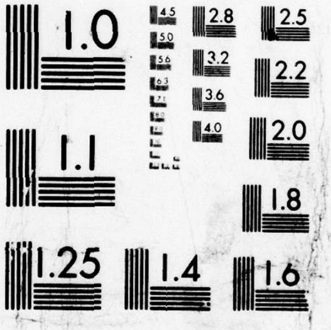
AD-A075 939 OFFICE OF NAVAL RESEARCH ARLINGTON VA F/G 15/5
NAVAL RESEARCH LOGISTICS QUARTERLY. VOLUME 26, NUMBER 3.(U)
SEP 79

UNCLASSIFIED NL

1 OF 2
AD
A075939



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

ER 1979
NO. 3
1278

ge
7
3
3
5
3
7
7

Vol. 26, No. 3

NAVAL RESEARCH LOGISTICS QUARTERLY

September 1979

6

NAVAL RESEARCH LOGISTICS QUARTERLY

Volume 26, Number 3:

11
SEP[redacted]979
VOL. 26, NO. 3

12 169

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

EDITORIAL BOARD

MANAGING EDITOR

Seymour M. Selig
Office of Naval Research
Arlington, Virginia 22217

ASSOCIATE EDITORS

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription Price: $11.15 a year in the U.S. and Canada, $13.95 elsewhere. Cost of individual issues may be obtained from the Superintendent of Documents.

The views and opinions expressed in this Journal are those of the authors and not necessarily those of the Office of Naval Research.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulations, P-35 (Revised 1-74).

PRICE #3.50 PER COPY

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

THE NEXT DECADE OF LOGISTICS RESEARCH*

Harvey M. Wagner

School of Business Administration
University of North Carolina
Chapel Hill, North Carolina
McKinsey and Co.
New York, N.Y.

ABSTRACT

Pathbreaking logistics research over the next 10 years will focus on systems problems. Whereas past research generally has taken a "bottom-up" approach, future investigations are likely to pursue a "top-down" philosophy. Specifically, attention will concentrate on diagnosis of systems' improvement potentials; easy-to-use analytic approaches, inherently approximative, will be devised for quickly ascertaining whether a complex operating system can be substantially and effectively improved. Theories to assist in overall systems design, particularly the setting of boundaries and buffers among systems components, will be developed. At the same time, techniques for accurately forecasting future systems performance will be investigated.

Underlying such research will be efforts to gain better understanding of management information requirements, including approaches for monitoring systems performance and providing early warning detection of systems degradation. Improved management information systems will have to be coupled with appropriate design of managerial organizations and assignment of decision making responsibilities. Important avenues of research will be development of robust approaches, that is, both mathematical techniques and organizational approaches that are not too adversely affected by limited data, a changing environment, and human frailty.

Finally, critical research will be directed at the implementation process, especially the interaction among initiation, design, testing, and ultimate adoption.

This prognosis will explore the above themes in the context of large-scale, complex systems. The decision areas will encompass inventory replenishment, multiechelon hierarchies for stockage and maintenance, procurement, transportation, scheduling, facilities planning, budgeting, reliability, and personnel management.

THE MOMENTUM OF HISTORY

Functional Subdivisions

The logistics functions in commercial and military organizations are so well established that their mission and performance often are taken for granted. Even when an organization

*This survey article was prepared for publication in *Science, Technology, and the Modern Navy*, E.I. Salkovitz, Ed. (ONR-37, 1977) in commemoration of the thirtieth anniversary of the Office of Naval Research. It is reprinted here because of its continuing timeliness and interest to the readers of the Naval Research Logistics Quarterly.

undergoes major structural renovation, the logistics functions may escape critical notice. Such activities traditionally are defined to include procurement (including purchasing of raw materials, packaging, product components, subassemblies, maintenance items, and capital equipment); manufacturing administrative processes (including scheduling of machinery, sequencing of work orders, selecting of manufacturing techniques); inventory control (including stocking of raw materials, in-process working inventory, and finished goods); distribution of resources that are held at various storage locations; and transportation (including selection of carriers, scheduling and loading of transportation equipment, negotiation of rates, and movement and deployment of personnel). In some organizations, logistics also encompasses maintenance and repair of equipment, reliability engineering, and facilities planning.

Despite the obvious connections among these functions, many organizations separate the responsibilities for the various logistics activities. As a result, the full economic and service improvement potential that could be realized by a coordinated effort is rarely achieved. Furthermore, logistics managers frequently are postured to have a reactive, rather than initiating, role. More specifically, logistics management is expected to execute requests from other parts of the enterprise, but not to actively suggest how overall integrative systems improvements can be made.

Today the costs of logistics have become sizeable, however, and subject to tighter managerial control, so that large organizations can no longer give short shrift to the logistics functions. To the contrary, many establishments have already made noteworthy improvements by eliminating trouble spots in their logistics functions. As we shall suggest, significant new opportunities can be created by an organization that recognizes and can thus coordinate the linkages among its various separate logistics functions.

Management Science Impact

Early in the evolution of management science and operations research, scientists realized that central logistics issues could be studied and eventually comprehended by means of the developing methods of applied mathematics. In particular, the researchers devoted a staggering amount of effort to formulating scientific models of inventory control; devising scheduling policies for equipment, projects, and production; using mathematical programming in planning analyses; testing operating doctrines for machine maintenance, repair, and replacement; evaluating options for transportation routing; and relieving congestion in queuing systems, to cite only a few of the classic problem areas.

The challenge of these problems has engaged the interest of talented scientists, including several recent Nobel Prize recipients. In addition to the intrinsic fascination of the problems' natural complexities, the research was impelled by the growing availability of large-scale electronic computers that presumably could perform numerous calculations and could store and process the data required to drive the model analyses to usable conclusions.

Without doubt, the degree of increased understanding afforded by the model building of management science and operations research in the past 30 years is impressive. An incredible amount of research has been done in fathoming the nature of logistics processes and their associated decisions, and there is no indication that interest and effort are waning.

Nevertheless, logistics managers are justified in questioning the extent to which the research findings have affected day-to-day decision making. Without denying that model-building research has brought significant systems improvements, such managers may express the wish that they could better use logistics models to help solve the remaining larger issues of the design and operation of entire logistics systems.

The Inward Spiral

As in all branches of applied science, an analytic problem, once defined, takes on a life of its own, regardless of its original source and setting. These problem situations seem to hold endless fascination for succeeding generations of scientists. The result frequently is a steady stream of refinements and extensions of the original formulation and analysis. These additions to knowledge may not be trivial from a technical point of view; their elegance and generality may warrant the intense intellectual effort spent producing them. Whether such progress helps solve the original real-life problem is another matter, however. The nature of model-building analysis is to abstract a piece of a complex problem that can be subjected to fruitful study. Unfortunately but inevitably, the resulting approximation to reality some times misses the target of providing a useful guide to decisionmaking. Ample evidence demonstrates that subsequent research often pushes the formative analysis further from reality—that is, makes progress in areas not pertinent to the critical limitations of the initial approximation.

Thus, despite the current active research in logistics processes, we cannot ensure that significant research breakthroughs will continue if we rely solely on letting past momentum determine the types of problems and the technical approaches of the future. To offset the natural tendency of applied research to spiral inward, logistics managers must energetically make known the problem areas that cry out for new analysis. Constant infusion of reality in logistics research is the best guarantee that the next decade of effort will have a major impact.

A SCORECARD OF RESEARCH PROGRESS

Bottom-up and Top-down Orientation

By and large, logistics models have focused on phenomena at the bottom levels of organizations. For example, the mathematical models derived over the past three decades have dealt with replenishment of individual stock items, initial provisioning of spare parts, sequencing of particular orders, overhaul of particular pieces of equipment, replacement of particular components, and so forth. A corollary is that these models have concentrated on single types of logistics decisions (replenishment, procurement, maintenance, transportation) rather than on systems of decisions. Even the notable exceptions to this generalization, such as in applications of mathematical programming models that deal with the deployment of limited resources, often treat as given certain assumptions that the highest level of management would prefer to consider as variables. To illustrate, in a transportation distribution study using mathematical programming, the analysis typically takes as given the products to be shipped and the customers to be served. Top management may be more interested in whether the products should be manufactured at all, whether certain customers are unprofitable because of the transportation rate structure, and how much service is required by customers. Of course, such issues can be sorted out in part with the aid of models, but in practice the typical study orientation has been to ignore such issues.

Another way of stating the point is to say that most management science and operations research models dealing with logistics have not begun by attacking the questions that would be posed by the topmost level of management. For example, when senior management is asked to approve a systems design effort to tighten inventory control, it wants an estimate of the savings potential of such a new design. When expansion of a factory warehouse is proposed, senior management wants an assessment of the possible share-of-market impact of having more or less stock at the location, which may be geographically removed from the company's customers. When a new product is to be introduced by a computer manufacturer, top management wants to know the economic ramifications of providing for concomitant repair and service, including the cost of parts replenishment. In brief, senior managements typically seek a comprehensive economic analysis of the "big picture."

Management scientists have assumed, almost as an axiom, that to obtain answers to high-level management questions, one must build the analysis from the bottom up. Thus, to predict an inventory system's performance, the researcher has been inclined to add up the performance characteristics of the individual components. Regretably, this bottom-up presumption has not proven itself to be without severe limitations. One difficulty has been the sheer effort involved in ascertaining and then "adding up" the component details. The analytic and data-processing difficulties that arise from starting at the bottom and aggregating up can be severe and can consume much of the analytic staff's time and energy. Ironically, in such instances senior management finds itself funding its own research project to learn whether the organization can benefit from previous logistics research.

To make matters worse, the "adding up" process may amplify rather than dampen the errors in the approximative assumptions of micromodels. When economies or diseconomies of scale, such as occur in the loading and routing of transport vehicles, are present, but virtually ignored by a microcosmic model, the consequent aggregation of individual calculations can be far off the mark. What appears to be an incidental approximation in the small can turn out to be a gross and misleading oversimplification in the large.

It is becoming clearer that these top management issues ought to be modeled in their own right. The potential advantages include faster and more accurate results. Even more important, perhaps, starting at the top affords a better opportunity to focus on issues, assumptions, and evaluation criteria that are most relevant to senior management.

So that there is no misunderstanding, we hasten to acknowledge that top-down analysis is not yet easy. In fact, we believe that this point of view will be a major focus of research over the next decade. The research tasks certainly will be at least as difficult and challenging as those that have been confronted with the bottom-up approach. Work to date suggests that considerable innovation will be required.

The Narrow End of the Time Tunnel

Logistics models have addressed management decisions that at one extreme pertain to daily phenomena, such as replenishment, scheduling, and repair, and at the other extreme, to long-range commitments, such as plant location, capacity expansion, and development of new products. A common observation is that at the first extreme the mathematical models are simpler to analyze (in the sense that they require less data and computation) but harder to implement (in the sense that they frequently require a sweeping systems design). In contrast, planning models for long-term decisions provide extremely useful information with a reasonable amount of effort, but involve an inordinately heavy use of computers and data manipulation.

Most logistics management functions in large enterprises involve an amalgam of both short- and long-term decisions. An important implication is that management of these enterprises must be prepared to deal with the different organizational stresses that arise from applying management science and operations research efforts at the two ends of the tim:-horizon spectrum. Research staff thus must include personnel capable of one-time innovative model building and data analysis as well as of designing and implementing operating systems.

Leashing the Crunchers

A curious paradox is connected with the use of large computers. As pointed out previously, advances in computer software and hardware technologies have spurred the development of logistics model building. It is inconceivable that the progress made so far in studying logistics decisions could have taken place if computer developments had leveled off. Furthermore, to the extent that such models have been applied to strategic as well as to operational decision-making situations, computers have been essential. Nevertheless, the difficulties in using computers in new model-building situations still are severe. In fact, even in so-called standard applications, such as the development of a new medium or large-scale linear-programming model, the tasks of collecting and analyzing the data, converting the data into model coefficients, obtaining usable optimization results, and providing management with readable analyses are now by no means routine. Admittedly, experienced technical experts now have a much better time of it than do novices. Also, today an organization receives considerably more "computation per buck" than it did a decade ago. Be that as it may, management must not view as insignificant the development and completion effort for a logistics model application. To add to the paradox, those software developments aimed at enhancing the application of a particular class of models, such as mathematical programming, have turned out to increase the learning setup time of beginners.

A related point is that, all of the statisticians' research not withstanding, model-building practitioners often are forced to resort to crude ad hoc data manipulation procedures in order to analyze historical information. Unfortunately, a model builder who has had a standard introduction to regression analysis, for example, is not very well equipped to detect, let alone design, useful data fitting formulas. Part of the difficulty, of course, is inadequate education. However, to offer a comparison, a logistics model builder need not be a highly trained technical expert or mathematician to run a standard linear-programming computer routine. Yet the same individual is almost certain to fail in manipulating a set of data on a dependent and several independent variables in trying to obtain a tight regression fit. (The usual approach is to employ standard multiple linear regression and hope that the resulting fit will be fairly good.) Oddly, most high-powered statistical routines now available on computers provide copious statistical tests that seem to make little sense to most users. Hence, data analysis for managerial decisionmaking is a burgeoning field with vast opportunities.

Management scientists and operations researchers are only beginning to come to grips with the intricate data analysis problems that arise in the use of computer simulations of stochastically driven systems. Of course, the complexity of such problems has been recognized for many years, but only recently has there been a better appreciation of how pervasive and knotty these difficulties are. The unsophisticated simulation model builder traditionally has assumed that all such estimation problems could be "bought off" by investing in a sufficiently long simulation history. In a trivial sense, that attitude is correct—but only lately has it become apparent that a sufficiently long history may be far longer than most practitioners would ever have guessed. Computation time is a scarce and costly resource, and the solution to these problems is not to run longer but to run smarter. At last this topic is under active research investigation.

Crossing the Technical Barriers

In the next section of this paper, we suggest several general classes of problems that will challenge future researchers of logistics decisions. Here we note a few of the technical problems that remain and attract the attention of researchers.

In one way or another, all realistic applications of model building to logistics decisions involve dealing with large-scale systems. The source of bigness may be the great detail that must be encompassed, for example, as in implementation of stockage rules for a system of tens of thousands of inventoried items, or the source may be the large number of options to be addressed, as in a multiperiod strategic planning model.

The problems of large-scale applications include both the sheer number of computations required as well as the vast amounts of input data that must be collected and reviewed and the resulting extensive output to be analyzed. Much progress is needed in techniques that help human analysts comprehend large sets of data. (Recent developments in computer graphics are good examples of what can be done to let a human literally see multidimensional phenomena.)

A related problem is the development of methods for testing model assumptions and data error sensitivity. Although many mathematical formulas have been developed to answer specific sensitivity questions about particular model structures (such as those that arise in analysis of linear-programming models), there is still no unifying approach or point of view for ferreting out which of the many parameters are most critical. A higher level of computer-assisted thinking is needed to alert the model builder to the weak points of the model.

Discontinuities, nonconvexities, and combinatorial phenomena are not yet completely under the thumbs of operations research analysts. Although significant progress has been made with such problems in the past 5 years, the halfway mark probably has not been reached.

Interestingly, the applied science community is not complaining that the mathematical problems are too complex to allow continued research progress. Progress seems slow, and the power required certainly is escalating, but there does not appear to be any din of discussion among management scientists and operations researchers centering on the few major unsolved technical problems that persist in defying successful attack. Rather, the lament is that problems currently under study are old-hat and of less intrinsic interest than those addressed in the early days of logistics research.

Without judging the validity or propriety of this lament, we argue in the next section that many important research tasks remain to be faced in the coming decade. As will be apparent from the discussion, the starting point for many of these topics is not the previously made generalization on the classic types of logistics models. Rather, the recommended approach is redefinition of the remaining problems, taking into explicit account the pressing needs of logistics managers. We propose a renewed and vigorous look at managers' topical problems rather than previous researchers' leftover problems.

THE CHALLENGES THAT AWAIT

A View to the Practical

In analytic research into logistics decisions, management scientists and operations researchers have been inclined to let the mathematical formulation of a model dictate or suggest the appropriate mode of analysis. For example, when decision problems have been posed in terms of dynamic-programming functional equations, then, generally, researchers have explored mathematical and computational ways to solve the functional equations. In inventory-control models, research has focused on ascertaining the form of an optimal policy and determining the computational implications of exploiting this knowledge of the optimal form. Similar illustrations could be cited for other types of probabilistic applications. Unfortunately, even after an initial mathematical formulation has been simplified by taking account of analytically derived information about the form of the model's solution, the complexity and the computational burden remaining is not trivial. As a result, applications of many such models have been limited, and sometimes even nonexistent.

An alternate approach, which is beginning to have some currency, is to derive simple but close analytic approximations to the original model. These approximations are easier to handle computationally and are therefore much more attractive from an applications point of view. (An example will be provided in the next section.) In most real-life situations the data required by a model are themselves approximate, by the very nature of their historical base. Hence, the degradation of economic performance due to analytic approximation may be negligible. Imperfect information typically overshadows the analytic approximation as a source of model error. Although numerical approximation is a seasoned topic in computer science and, to an extent, in statistics (by way of curve fitting), the subject is relatively new in operations research. It offers considerable promise and may make practical the solution of many models that have been discarded earlier as computationally unwieldy.

A related technique is to derive analytic models with parameter values that are numerically fit from a limited discrete set of optimal points (policies). These fitted relations permit interpolation of intermediate parameter settings. In other words, the researcher starts with a grid of parameter values, performs the detailed model optimizations to derive the best policies for the grid, and then fits an analytic function of the parameter values to the set of numerical policies.

A similar vein of research is to discover the actual sensitivity of optimal policies to various parameters of a model. Evidence is building that many models that appear to involve multivariate optimization can without much loss be factored into separate optimizations, each requiring an easier manipulation of fewer variables.

In summary, considerable future research will be turned to investigating the numerical properties of logistics models, with emphasis on parameter settings that are relevant for actual applications. Such investigations will result in computational models that are simpler to use and thus will enhance the applicability of the models.

Breakdown of the Boundaries

Perhaps the most important of all the new avenues for future research will be modeling efforts that combine heretofore separate investigations of logistics decisions. Examples abound in military logistics systems. There are, for example, significant economic tradeoffs relating to initial procurement, spares provisioning, location of repair facilities, design of component parts, and installation of data collection systems to track weapon-system performance. Similar illustrations are easily cited in commercial organizations. For example, a manufacturing company must balance off considerations of labor stability, the buildup of seasonal inventories, the location of such inventories, the mode of transportation to customers, the frequency of delivery in relation to the capacities of transport vehicles, and the targeted service performance (that is, availability of stocks and promptness of delivery).

A bottom-up approach for investigating the interactions among logistics functions does not seem as promising or as practical as a top-down approach. In constructing a top-down model, however, a researcher should keep in mind the operating characteristics of low-level logistics models and include these characteristics in the formulation of the high-level model. For example, if a segment of an inventory system has a square-root relational dependency on the annual demand for the encompassed items, then that system's numerical phenomena should be included in the model specification.

Because of the inherent complexity of multifunction models, a successful analytic approach may involve exploring only a set of case studies rather than seeking some sort of global, or even local, optimum. In other words, the model builder may have better success in investigating plausible solutions and, with feedback, refined versions of the alternatives, than in trying to simplify the interconnections in the mathematical structure to permit "automatic" optimization algorithms. The case-study approach to integrative analyses also facilitates the inclusion of discontinuous economic and physical phenomena. After the number of high-level decision options has been narrowed to a select and attractive few, then the now-familiar lower level model-building approaches can be brought into play to refine the analyses if need be.

The Human Side of Systems Design

It is surprising, perhaps shocking, that virtually no research attention has been given to the human factors aspect of modern logistics systems design. If logistics research is to become part of the warp and woof of an organization, attention must be given to the organizational setting, including the assignment of responsibilities. For example, even if model builders succeed in breaking down the boundaries between logistics functions, little benefit will result if there is no corresponding integration of management logistics responsibilities. In a manufacturing company, the links between sales forecasting, production planning, and materials purchasing are critical to the economic functioning of each of these activities. A comprehensive logistics model would combine the three elements, but the model would not produce results unless the three functions were controlled by a consistent corporate-wide logistics management policy.

The organization of most logistics operations in an enterprise is based on historical evolution; changes have taken place, if at all, typically at times of crisis. Yet almost always large improvements can be made as a result of a comprehensive look at the logistics needs of the organization. More often than not, much of the improvement devolves from realignment of responsibilities along with appropriate management review and control, rather than from revision of isolated decisionmaking processes, such as production scheduling. In other words, most separate logistics functions fare pretty well given the organizational constraints under which they operate; any noteworthy improvement comes from breaking down some of the constraints.

Considerable future research effort is required not only in thinking through organizational structure, but in examining effective approaches to personnel motivation, the communication of information for decisionmaking, and management review and control, insofar as these human activities bear on the design of integrative logistics systems. Logistics personnel in most enterprises are prone to a "beat-the-system" attitude; this proclivity should be recognized explicitly and factored into the systems design process.

Finally, even assuming benign attitudes within an organization, researchers must explore ways to improve the interactions between personnel (managerial, staff, and clerical) on the one hand and computer-driven data systems on the other. The notion that a computerized logistics system is conducive to easier decisionmaking is too naive to be of value. In fact, a computerized approach often seems to make some jobs harder and others duller. Rarely does the implementation of such a system result in an upgrading and simplification of jobs throughout. The commonly expressed negative attitudes about computer systems in large organizations are grounded in considerable experience, and the root causes call for careful study.

A Window on the Future

Now that 30 years of logistics research have passed, senior-level management has come to feel that it should be possible to diagnose the need for systems improvement without undertaking a major, lengthy research project. It is incredible to such managers that systems analysts are unable after a brief investigation to at least scope out a reasonable range of improvement potential from contemplated systems revisions. But strange as it may be, management scientists and operations researchers have made little progress in devising powerful diagnostic tools. That should be given priority. The effort will have to be empirically based in part, at least insofar as the suggested approaches should stand the test of actual field validation. The purpose of these diagnostic tools is to provide management with estimates of the future benefits of a commitment to invest in systems revision. A top-down orientation would seem to provide the proper perspective.

A similar, possibly more technical topic is study of methods for predicting systems performance when new decision rules are to be used. In this context, suppose that a proposed design has been worked out in detail, but that some of the parameter settings used in the design remain under investigation. As an example, perhaps the frequency of data revision and file update is in question. Systems performance characteristics often are investigated by means of a simulation. Such simulations usually are computer models themselves, but sometimes, especially in military systems, they are onsite tests. Little scientific research has been done to establish the validity of these predictive approaches. Practical considerations frequently rule out routine application of classical statistical design-of-experiments methods. In the methods commonly used in practice, often a bias exists that makes a proposed system design appear to perform better than it will in fact. The source of the bias is easy to detect, once one is alert to its possible existence, but correcting it may be difficult. In admittedly over-simplified terms, the bias arises because the new design itself has been fashioned according to historical data, and therefore it appears to perform well in historical perspective. The inescapable difficulty is that of necessity many models are driven by historical information that may be so limited as to prohibit using a "split-sample" approach to validation.

A related need is for monitoring devices and early warning controls that automatically determine when a new systems design revision may be warranted. Presumably, if progress is made in fashioning diagnostic and predictive tools, the way will be paved for the devising of continuing controls that automatically determine when a new systems design revision may be warranted. Here too, a top-down approach seems appropriate. It may be very difficult to detect any systems' performance degradation by looking at individual components one by one. Sensitive aggregates, if such can be found, are needed.

Disaster Insurance

Mathematical programmers have learned an important lesson that should be noted by all model builders. A single-criterion optimization model typically pushes to the greatest extent possible each simplifying assumption in a model. For example, if a nonlinearity has been approximated, the optimization process will find how to exploit the *approximation*. As a result, the solution may strain the assumptions beyond credibility and usability.

To the extent that logisitcs research model building will break down the barriers between functions, as proposed earlier, care will have to be taken that the resulting solutions are not "too tightly tuned." The organization must be able easily to buffer unexpected (unmodelled) events. It is likely that second-best (less-than-first-best) strategies may be preferred if they do not force the organization into assuming a confining posture. Observers of real organizations recognize that most managements, usually with good reasons, shy away from strategies that have serious downside risks. Aside from recognizing the existence of multicriteria problems, management scientists and operations researchers have not made much progress in discovering the sensitivity of strategies to criteria that recognize and avoid downside risks.

The goal-establishment problem is not solely technical; it also concerns the organizational issues mentioned above. The enterprise must build in buffers, by a careful structuring of the organization, to absorb unplanned-for shocks. To illustrate, the production management component of a system may need to have a backlog of maintenance projects to fill up slack time that may arise when the marketing organization has been over-optimistic in its forecasts of sales.

To the extent that approximate models will be devised, care must be taken that the recommended decisions do not degrade too badly when the model's assumptions become invalid. For example, even though there may be very little lost in the original optimization model when a parameter is misspecified, the same need not be true in the approximative version. The chief source of misspecification in real applications is the uncertainty about future demand, failure rates, procurement costs, transport reliability, and so forth.

Getting the Job Done

The process of systems implementation deserves attention in its own right. It has become apparent that the full process of implementation has many components, some of which concern the nature of the decision problem, some the organizational setting, and some the support systems design. It is important that a framework of analysis be established to piece together the essential components, namely the decisions affected, the targeted benefits, the downside risks, the assignment of responsibilities, the development of the systems approach, the education of managers and support staff, the inherent life cycle of the application, the specific systems design, the required data, and the model's validation.

In addition, it would be helpful to examine managers' psychology with regard to systems' development authorization—for example, how do they view associated career development hazards, assess the reasonableness of a project's timetable, decide whether the design will be useful and avoid being embarrassed by an unsuccessful outcome.

The proper methodology for studying implementation is itself a research issue. The term "implementation" actually presents a problem of definition and, in any event, implies a value connotation in that agreeing to implement is normally presumed to be good and failing to implement to be bad. To make sense out of implementation processes, researchers must establish standards of comparison that are legitimate within a single organization as well as across organizations.

Summary

This section has touched on a number of avenues of research in logistics systems design that could have significant impact if successfully pursued. In looking back over the list, it is clear that the suggestions are not aimed at particular types of logistics decisions. They are aimed, rather, at a type of approach that cuts across individual logistics decision areas. Hopefully, the list makes clear those challenges that stem from recognition of organizational and managerial needs in relation to unsolved and mind-boggling technical puzzles. Assuredly, the suggested research areas are replete with tough analytic tasks, and the technical inspiration required will not derive solely or even mainly from the methods of past applied logisitics research.

A GLIMPSE AT THE POSSIBLE

Strategy for Research

A rich variety of applied mathematics approaches has become standard in management science and operations research studies of logistics processes. They include mathematical programming optimization, dynamic programming, Markovian analysis, and computer simulation, to name only the more prominent. The primary role of computers has been to perform algorithmic computations on particularized versions of mathematical programming models and to provide simulated results for (typically) stochastic systems run with special settings of the underlying model's parameters.

Intererstingly, the computer has seldom been used to ferret out the *qualitative* properties of models, to provide the analog of the physical scientist's experimental laboratory. We believe that substantial breakthroughs are possible in many logistics research problems that are now deemed intractible because the standard applied mathematical approaches have been pushed to their limit. We suggest and illustrate in this section how computers can be used to provide new analytic models capable of solving some currently unanswered high-level management questions.

A Case in Points

Take as an example the subject of inventory control. Over the past two decades, mathematical analysis of inventory stockage models has made great progress, and real-life implementation of inventory systems, based at least in part on the results of this modern research, has taken place. Nevertheless, when an organization considers the possibility of designing and installing a new replenishment system, senior management typically finds it arduous and time-consuming to obtain reliable answers to questions such as

- What are the effects of consolidating demands from several different warehouses into a single central warehouse?
- If system-wide demand increases (through, for example, an enlarged share of the market), what are the resulting cost and service implications?
- How much is it worth to obtain quicker delivery of replenishment orders?
- By how much will costs rise if service is increased?
- How will costs be affected by less frequent updating of information?

For some of these questions, no easy-to-use analytic formulas have been devised. For others, an answer is forthcoming only if the analyst painstakingly uses a bottom-up approach, that is, makes the calculations for each of a number of individual stockage items and then aggregates the results.

Recently an alternative analytic approach has been investigated by the author and his associates, Richard Ehrhardt, Alastair MacCormick, Ronald Kaufman, Arthur Estey, and John Klincewicz. A capsule view is provided below to indicate the nature of the research strategy.

Systems Design Scenario

Consider an inventory manager who must design a system of replenishment rules for the stockage of possibly thousands of items. Assume that the manager can specify a criterion function to determine whether one system design is better than another. Suppose that the manager has elected to use so-called (s,S) policies: when inventory on hand and on order falls below s, place an order so that, as a consequence, inventory on hand and on order equals S. It is necessary to compute numerical values for the pair (s,S) for each item to be stocked. Under widely applicable conditions, it is possible to employ an algorithmic approach that provides optimal values for (s,S), but the computations are numerous and make application to a large-scale system prohibitive. Further, the optimizing algorithm assumes that the demand distribution for each item is known exactly; this is virtually never true in practice. The manager inevitably must use past data to *estimate* the demand distribution.

The systems designer's tasks then include selecting in concert the number of historical observations to use, the frequency for repeating the reestimation process, the form of the replenishment rule, the statistical estimators to produce the demand parameters required by the rule, and the design parameters of the rule, namely, the values of s and S in our illustration. Typically the manager makes all of these choices, at least in part, according to simulations of how the proposed system would have performed in the past. In doing so, the manager typically uses the same limited data for both estimating the demand parameters and predicting systems performance.

Recognizing and Attacking the Issues

Eventually, inventory managers will have to provide the answers to the questions posed by senior management. But even before attacking top management's questions, the designer must find a practical approach to the mundane issues of calculating the rule values themselves and discovering how accurate the retrospective predictions are likely to be. Regretably, these tasks are mathematically so complex that they do not appear tractible by known methods of applied analysis.

It is possible to make considerable headway, however, by devising an experimental design approach with the further help of a computer, first postulating a set of parameter values that encompasses most of the cases likely to be encountered. For the sake of definiteness, suppose that the parameter values are given as in Table 1.

We examine a full-factorial representation of all levels of these parameters in combination with each other, yielding a total of 288 settings. Using exact computations, we find the corresponding 288 optimal (s,S) policies. Next, using standard curve-fitting techniques on these 288 pairs (s,S), we obtain numerical approximations for the quantities $D=S-s$ and s. Specifically, according to Ehrhardt, we obtain the formulas

$$D = (1.463)\mu^{.364}\,(K/h)^{.498}\times$$
$$[(L+1)\sigma^2]^{.0691}$$

and

$$s=(L+1)\mu + [(L+1)\mu]^{.416}\times$$
$$(\sigma^2/\mu)^{.603}\,U(z),$$

where

$$U(z) = .182/z + 1.142-3.466z$$

$$z=\left\{\frac{\mu^{.364}(K/h)^{.498}}{(1+\frac{p}{h})\,[(L+1)\sigma^2]^{.431}}\right\}^{1/2}$$

TABLE 1 – *System Parameters*

Factor	*Levels*	*Number of Levels*
Demand Distribution	Poisson $(\sigma^2/\mu=1)$ Negative Binomial $(\sigma^2/\mu=3)$ Negative Binomial $(\sigma^2/\mu=9)$	3
Mean Demand μ	2, 4, 8, 16	4
Replenishment Leadtime L	0, 2, 4	3
Replenishment Setup Cost K	32, 64	2
Unit Penalty Cost p	4, 9, 24, 99	4
Unit Holding Cost h	1	1

To test whether this approximation is close enough (near optimal), we derive the 288 approximate (s,S) pairs, calculate their corresponding expected cost using *exact* formulas, and compare the associated cost with the original optimal cost. In this design, 95% of the 288 cases are within 1% of optimal. Then we examine the robustness of the approximation by trying a number of interpolated and extrapolated sets of parameter values. (In such tests, we had equally good results.)

Thus the curve-fitting exercise provides the systems's designer with an easily computed replenishment rule that depends on the economic parameters and only the mean and variance of demand. But since the mean and variance are not known in real-life applications, the next step is to ascertain how well the approximation works in a statistical environment.

Presumably, in an actual situation the mean and variance of demand for each item would be estimated by the usual statistical techniques, that is, by computing a sample mean and variance from a limited history of data, and substituting these values into the approximation formulas. Again for the sake of definiteness, suppose that the designer wishes to investigate three possibilities: updating s and S (by recomputing the historical mean and variance of demand) every 13 weeks, or every 26 weeks, or every 52 weeks.

We can test the performance of the approximation rule under these different circumstances by running a computer simulation for each possibility. In particular, we again can choose a factorial design for the parameter settings, simulate the use of the rule for a sufficiently long history and for each of the three revision possibilities, and at the same time simulate the retrospective approach to predicting the future performance of the rule. In summary, we found that systems costs increase, on the average, by 20% above the optimal with complete information when only 13 weeks of data are used and variance/mean = 9; by 11.5% when 26 weeks are used; and by 6.3% when 52 weeks are used. For these same three cases, the forecast of systems cost performance are, respectively, 25.1%, 17.1%, and 10.7% *under* the actual values; interestingly though, most of the underestimation comes from the service (stockout cost) component, and the separate predictions of inventory and replenishment costs are typically less than 2% under the actual values.

Finding Systems Response Functions

Next we are ready to obtain simple-to-use analytic expressions for the total costs of using the approximate policies. We again employ for this purpose a curve-fitting approach. For the situation in which the mean and variance can be exactly specified, we derive

$$\text{Total Cost} \cong 5.663 h \mu^{.4495} (L+1)^{.2528} (p/h)^{-.9230/p\, +\, .1357} (K/h)^{.203},$$

assuming that variance/mean = 9. Similarly, when 26 weeks of data are used to estimate the mean and variance, we find

$$\text{Total Cost} \cong 3.798 h \mu^{.4309} (L+1)^{.3024} (p/h)^{.2550} (K/h)^{.1917}.$$

These cost functions provide the needed wedge into the problem of answering senior management's questions about forecasts. To illustrate, if mean demand doubles, total cost will increase by 20% in both cases. If, for example, the demands from eight independent and identical warehouses are consolidated in a single central warehouse, total cost will be reduced by about 68% in both cases. If service protection is increased from 0.9 in-stock probability to 0.95, it can be demonstrated that total cost will rise by 25% in the statistical environment. If leadtime is cut in half at the expense of doubling setup cost, then total cost in a statistical environment is reduced by 7% (*after* the higher setup costs are paid). If the system is updated only half as often, total costs may be reduced substantially; for example, if inventory costs are charged on end-of-the-review-period levels (as is frequently done for the property tax valuation component), the cost reduction is near 40%.

The above discussion has focused on total costs, but similar systems-wide approximations have been derived for each of the components of total cost and other operating characteristics.

Summary

What this abbreviated survey of recent inventory research advances has demonstrated is the way in which seemingly intractible mathematical problems can be solved by empirical and statistical investigation. Like any experimental approach, the suggested research strategy requires careful prior planning and sufficient completion time. The impressive tightness of the approximations, however, is encouraging.

EXPECTATIONS FOR THE FUTURE

Perceiving the Sector Factor

Unquestionably there are important differences between the private and public sectors in solving real logistics problems. The obvious differences are related to the sheer possibility of truly integrating separate logistics functions, the limited budgetary and personnel resources for systems redesign, and the fiscal constraints on any implied multiyear spending. Beyond these are differences in the basic missions of the logistics function. In a commercial enterprise, the logistics decisions support the buying, making, and selling functions and rather clearly lead to an eventual profit-and-loss impact. But in a military environment, the logistics mission is highly intertwined with the critical notion of combat readiness, which in the final analysis is only rarely tested and then under crisis circumstances. Perhaps ironically, it is in a military setting that the top-down approach to logistics is most essential, because very large sums of dollars are committed by the logistics decisions, and these must be balanced off against dollars spent on other military readiness functions.

Watching the Sign Posts

A truly telltale criticism of past management science and operations research investigations into logistics functions is that they rarely reflect timely economic issues. To illustrate, one is hard pressed to find in the applied-mathematics-oriented logistics research literature a careful discussion of the impact of inflation, the limited availability of fuels and other strategic resources, or the rate of technological change. However, actual logistics managers are painfully aware of these environmental changes and their impact on logistics decisions. Logistics research will only stay vital if it pays heed to the changing world.

Generating Viable Options

It is virtually a tautology to say that a formal logistics decision model encompasses a static universe of options. The solution drawn from this universe by the model may or may not yield a recommendation that can be implemented, but if the solution is unacceptable the analyst always can go back to the drawing board, revise the model, and try again. What is more important to the search for significant progress in logistics decisionmaking is to concentrate on discovering truly new options. Without sinking into a philosophical quagmire of subtle distinctions, we suggest that analysts pay more attention to relieving constraints, finding new conceptions and criteria, combining separate processes, and so forth than to searching for the very best answer within a well-established framework of concepts, laid down constraints, and circumscribed functions.

Substitution at the Margins

A related topic is the necessity that a wide view be taken of the important substitution possibilities. For example, there are tradeoffs between computer information systems and skilled labor, between large stocks of disposable spares and limited stocks of high-technology components, between fast modes of transport and extensive amounts of inventory, and between rapid communications systems and multiple pipelines, to name a few. The point is so obvious that it may not seem worth making, except that most logistics research takes place in a very limited context. The analyst may be either proscribed from examining such tradeoffs or ignorant of their existence and feasibility. Thus, one function of senior management is to encourage logistics staffs not to be too circumscribed in considering possibilities. An ancillary observation is that a logistics organization making such investigations must have access to a broad spectrum of skills and knowledge.

Next Up

In summary, this survey has attempted to realistically assess both the strengths and the limitations of logistics research to date and to generate excitement and enthusiasm for the worthwhile but difficult tasks ahead. Our prognosis is that substantial advancements will be made in the coming decade by researchers who focus on problems at the traditional boundaries of the logistics functions, who keep abreast of the changing outside environment, and who break away from sole reliance on the well-worn applied mathemathics techniques that have already run their courses with regard to many now-classic logistics problems. None of our exhortations is meant to detract, however, from the unassailable value of building on past research momentum. We have tried, rather, to indicate where we think some of the still-buried great treasures are to be found in the next 10 years of logistics research.

GEOMETRY OF THE TOTAL TIME ON TEST TRANSFORM*

Richard E. Barlow

Department of Industrial Engineering
and Operations Research
University of California, Berkeley
Berkeley, California

ABSTRACT

Total time on test (TTT) plots provide a useful graphical method for tentative identification of failure distribution models. Identification is based on properties of the TTT transform. New properties of the TTT transform distribution are obtained. These results are useful to the user of TTT plots. Although IFR (DFR) distributions are particularly easy to identify from TTT plots, the user must exercise caution relative to identification of IFRA (DFRA) distributions.

1. INTRODUCTION

The geometry of the total time on test transform is helpful in interpreting total time on test data plots [1]. In particular, it is possible to infer tentative probabilty distribution models based on total time on test plots.

Let F be a failure distribution, i.e., $F(0^-) = 0$ and $\bar{F} = 1 - F$. Define

$$H_F^{-1}(t) = \int_0^{F^{-1}(t)} \bar{F}(x)\,dx \qquad 0 \leqslant t \leqslant 1 ,$$

the *total time on test transform* of F. It is easy to verify that, H_F, the inverse of H_F^{-1} is a distribution function. Also, H has support in $[0, \theta]$ if θ is the mean of F, since

$$\int_0^{F^{-1}(1)} \bar{F}(x)\,dx = \int_0^{\infty} x\,dF(x) = \theta$$

by an integration by parts. It is easy to verify that if $F(x) = 1 - e^{-x/\theta}$, then the corresponding $H_F(x) = x/\theta$ for $0 \leqslant x \leqslant \theta$. The result that our transform carries the exponential distribution into the rectangular distribution on $[0, \theta]$ is important.

As was proved in Barlow, Bartholomew, Bremner and Brunk [2], total time on test data plots tend to the total time on test transform of the underlying failure distribution as the sample size tends to infinity. In order to interpret total on test data plots, we need to understand the relationship between F and its transform. The following table summarizes the connections.

*This research has been partially supported by the Office of Naval Research under Contract N00014-75-C-7081 and the Air Force Office of Scientific Research (AFSC), USAF, under Grant AFOSR-77-3179 with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

This is part of a chapter in the forthcoming volume *Inference and Data Analysis for Reliability and Life Testing* with Frank Proschan.

TABLE 1 — *Logical Connections Between Life Distributions, Hazard Functions and TTT Transform Distributions*

Life Distribution F		Hazard Function $R = -\log \bar{F}$		Total Time on Test Transform Distribution H_F
Exponential	$\Leftrightarrow$	linear	$\Leftrightarrow$	linear
IFR	$\Leftrightarrow$	convex	$\Leftrightarrow$	convex
DFR	$\Leftrightarrow$	concave	$\Leftrightarrow$	concave
IFRA	$\Leftrightarrow$	starshaped	$\Rightarrow$	starshaped
DFRA	$\Leftrightarrow$	anti-starshaped	$\Rightarrow$	anti-starshaped

A function g defined on $[0,b)$ such that $\frac{g(x)}{x}$ is nondecreasing on $[0,b)$ is said to be starshaped with respect to the origin. If $G(x) = 1 - e^{-x}$, then F is IFRA (for increasing failure rate average) if and only if $\frac{G^{-1}F(x)}{x}$ is nondecreasing for $0 \leqslant x \leqslant F^{-1}(1)$. The function $G^{-1}F(x) = R(x)$ is said to be starshaped with respect to the origin. As the last two implications indicate, IFRA and DFRA distribution families are *not* characterized by corresponding properties of the TTT transform distribution. However IFR and DFR distribution families are characterized by corresponding properties of the TTT transform distribution.

To verify the implications in the table for the IFR (DFR) case, first assume F absolutely continuous with failure rate function, r. If F is IFR (DFR), then

$$\frac{d}{dt} H_F^{-1}(t)\Big|_{t=F(x)} = \frac{1}{r(x)}$$

is decreasing (increasing) in x which implies H_F^{-1} is concave (convex), i.e., H_F is convex (concave). Conversely, if H_F^{-1} is concave (convex), the failure rate function is increasing (decreasing). To see this, note that every IFR (DFR) distribution can be approximated arbitrarily closely by an absolutely continuous IFR (DFR) distribution. Since the limit of a sequence of concave (convex) transforms is concave (convex) on $[0,1]$, it follows that F is IFR (DFR) if and only if H_F^{-1} is concave (convex).

The IFRA distributions govern the lifelength of coherent systems with statistically independent components whose life distributions are IFR (or, more generally, IFRA). (Birnbaum, Esary and Marshall [3] or Barlow and Proschan [4]. They also arise in other reliability contexts. For these reasons, we are interested in the transforms of IFRA distributions. In the next section we show that if F is IFRA, then its transform distribution, H_F, is starshaped; i.e., $\frac{H_F(x)}{x}$ is nondecreasing in $0 \leqslant x \leqslant \theta$. Unfortunately, the converse is not true.

Bergman [5] and Mark Brown (personal communication) have pointed out that F is New Better than Used in Expectation (NBUE); i.e.

$$\int_x^\infty \frac{\bar{F}(u)}{\bar{F}(x)}\, du \leqslant \mu \quad x \geqslant 0$$

if and only if

$$H_F^{-1}(t) = \int_0^{F^{-1}(t)} \frac{\bar{F}(u)}{\mu}\, du \geqslant t \quad 0 \leqslant t \leqslant 1$$

where μ is the mean of F. (To see this, let $t = F(x)$ and make a change of variable using the fact that $\mu = \int_0^\infty \bar{F}(u)\,du$.) The NBUE class properly contains the IFRA class. Hence if the TTT plot is anti-starshaped, it necessarily lies above the 45° line and this is evidence that F is at least NBUE if not IFRA.

2. PRESERVATION OF PARTIAL ORDERINGS ON CLASSES OF FAILURE DISTRIBUTIONS

Let $R(x) = -log\bar{F}(x)$ be the hazard function of F as before and let $G(x) = 1 - e^{-x}$. Observe that $G^{-1}F(x) = R(x)$ so that if F is *IFR*, $G^{-1}F(x)$ is convex on the support of F and conversely. If F is *IFRA*, $\frac{G^{-1}F(x)}{x}$ is nondecreasing in $x \geqslant 0$ and conversely. This leads to a partial ordering on the space of failure distributions which we call "star ordering." Let $\mathcal{F}$ be the class of continuous distributions on $[\,0, \infty)$ and {deg.}, the class of degenerate distributions.

DEFINITION:

$F_1 \underset{*}{<} F_2$,(i.e., is star ordered with respect to F_2 if $F_1, F_2 \in \mathcal{F} \cup$ {deg.} and $\frac{F_2^{-1}F_1(x)}{x}$ is nondecreasing in x for $0 \leqslant x \leqslant F_1^{-1}(1)$).

According to this definition, every distribution in $\mathcal{F}$ is star ordered with respect to a degenerate distribution. Let $F_\alpha\,(x) = 1 - e^{-x^\alpha}$ for $x \geqslant 0$ and $\alpha > 0$. It is easy to show that if $0 < \alpha_1 < \alpha_2$, then $F_{\alpha_2} \underset{*}{<} F_{\alpha_1}$. Since F_α has failure rate $\alpha x^{\alpha - 1}$, it is clear that the failure rate of F_{α_2} is "increasing faster" than the failure rate of F_{α_1}. If $0 < \alpha < 1$, F_α is DFR. If $\alpha > 1$, F_α is IFR and F_1 is exponential.

DEFINITION:

$F_1 \underset{c}{<} F_2$ (i.e., F_1 is convex ordered with respect to F_2 if $F_1, F_2 \in \mathcal{F} \cup$ {deg.} and $F_2^{-1}\ F_1(x)$ is convex in x for $0 \leqslant x \leqslant F_1^{-1}(1)$).

It is not hard to show that c-ordering implies star ordering, but not conversely. Our main theorem is that the *TTT* transform distribution preserves both orderings.

THEOREM 2.1:

Let $F_1, F_2 \in \mathcal{F}$.

(a) If $F_1 \underset{c}{<} F_2$, then $H_{F_1} \underset{c}{<} H_{F_2}$.

(b) If $F_1 \underset{*}{<} F_2$, then $H_{F_1} \underset{*}{<} H_{F_2}$.

The following corollary provides the primary application of the theorem.

COROLLARY 2.2:

If $F_1, F_2 \in \mathcal{F}$ and $F_1 \underset{*}{<} F_2$, then

(a) $\dfrac{H_{F_2}^{-1}(t)}{H_{F_1}^{-1}(t)}$ is nondecreasing in $0 \leqslant t \leqslant 1$;

(b) $\dfrac{H_{F_1}^{-1}(t)}{H_{F_1}^{-1}(1)} \geqslant \dfrac{H_{F_2}^{-1}(t)}{H_{F_2}^{-1}(1)}$ for $0 \leqslant t \leqslant 1$.

PROOF OF COROLLARY:

By Theorem 2.1, Part (b), $H_{F_1} <_* H_{F_2}$ so $\dfrac{H_{F_2^{-1}} H_{F_1}(x)}{x}$ is nondecreasing in $0 \leqslant x \leqslant F_1^{-1}(1)$. Let $t = H_{F_1}(x)$ and Part (a) of the corollary is immediate. (b) is a trivial consequence of (a). | |

Figures 1 and 2 are graphical plots of the scaled transforms of gamma and Weibull distributions. They visually confirm Part (b) of the corollary. Figure 3 shows the same ordering as in Figures 1 and 2 with respect to the shape parameter of a lognormal distribution although this distribution is neither IFRA nor DFRA.

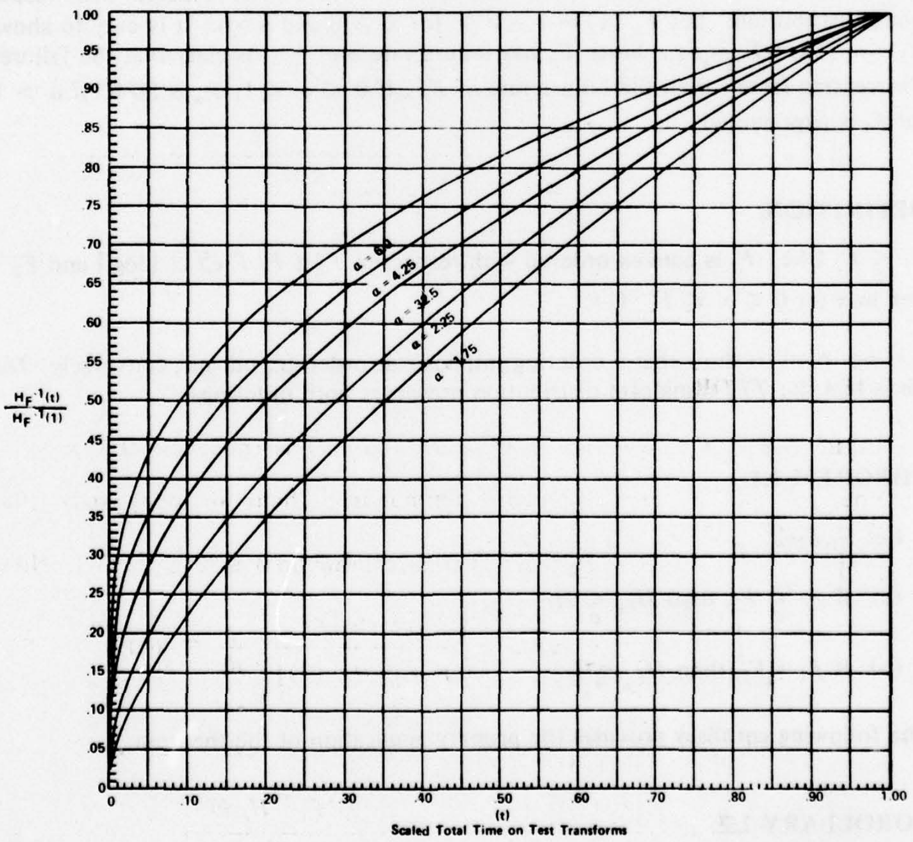


FIGURE 1. Gamma distribution (shape parameter α)

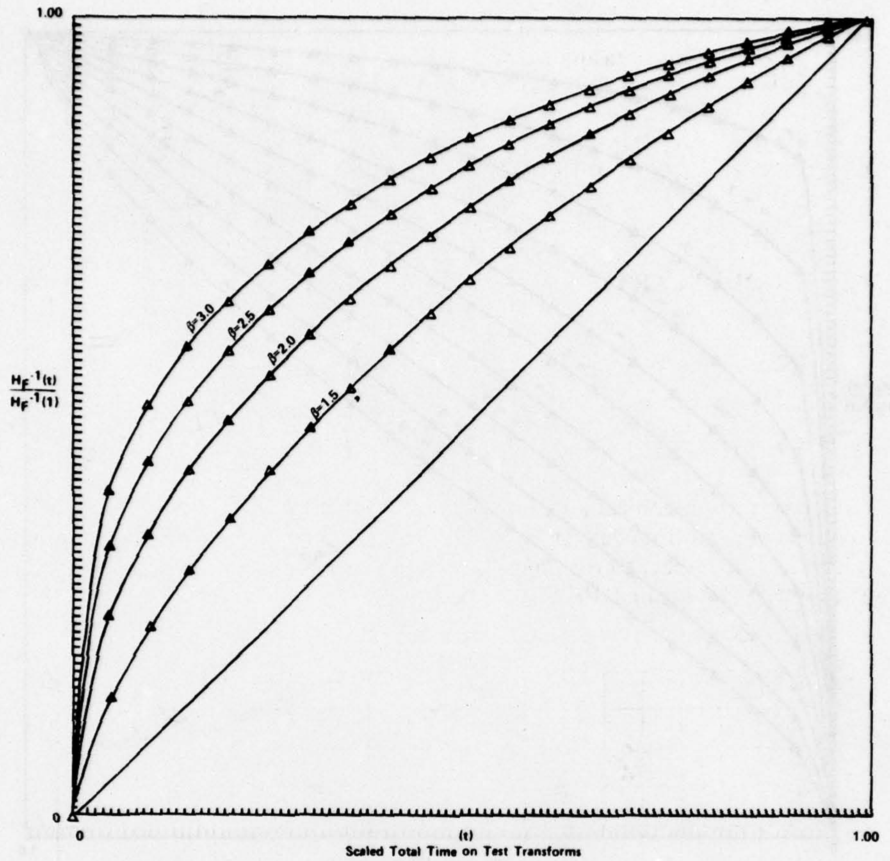


FIGURE 2. Weibull distribution (shape parameter β)

PROOF OF THEOREM 2.1:

(a) Assume $F_1 \underset{c}{<} F_2$. We wish to show $H_{F_2}^{-1}\, H_{F_1}(x)$ is convex in $0 \leqslant x \leqslant F_1{}^{-1}(1)$. First assume F_1 and F_2 are absolutely continuous. Then we need only show $\frac{d}{dx}$ $H_{F_2}^{-1}\left[H_{F_1}(x)\right] = \frac{d}{dx} \int_0^{F_2^{-1}[H_{F_1}(x)]} \bar{F}_2(u)\, du$ is nondecreasing in $0 \leqslant x \leqslant F_1^{-1}(1)$. Now

$$\frac{d}{dx} \int_0^{F_2^{-1}[H_{F_1}(x)]} \bar{F}_2(u)\, du = \left[\frac{1 - H_{F_1}(x)}{f_2\left[F_2^{-1}\; H_{F_1}(x)\right]}\right] \frac{d}{dx}\, H_{F_1}(x).$$

Let $x = H_{F_1}^{-1}(t)$ so that $\frac{dx}{dt} = \frac{1-t}{f_1[F_1^{-1}(t)]}$ and

$$\frac{dt}{dx} = \frac{f_1[F_1^{-1}(t)]}{1-t}\Bigg|_{t = H_{F_1}(x)} = \frac{f_1[F_1^{-1}\; H_{F_1}(x)]}{1 - H_{F_1(x)}}$$

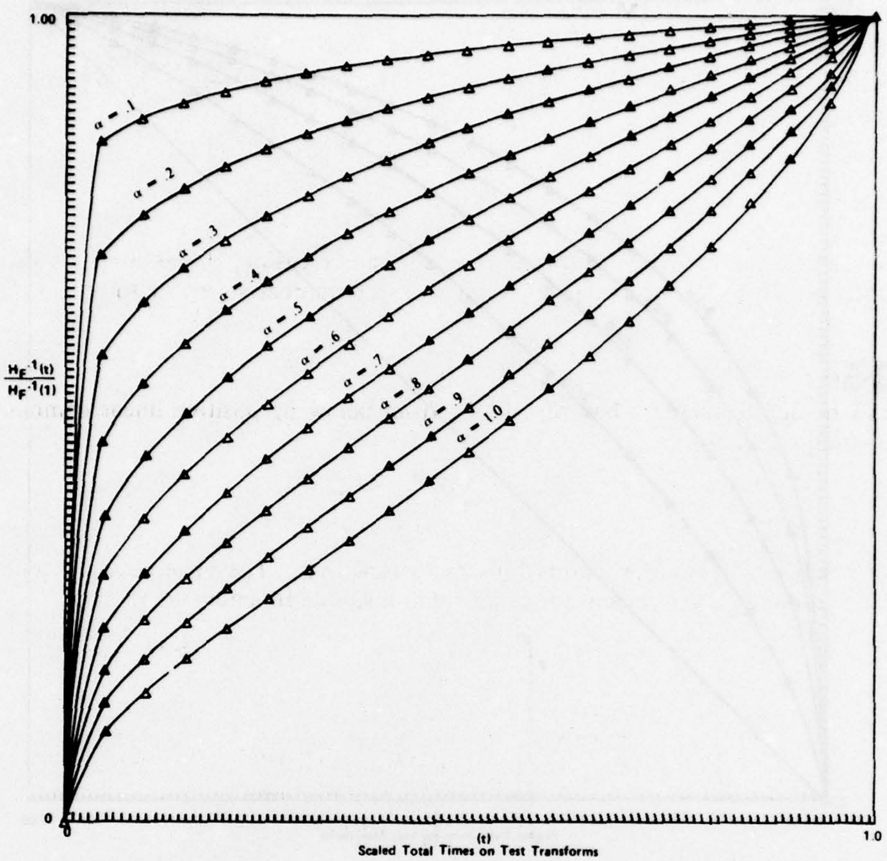


FIGURE 3.. Lognormal distribution (shape parameter σ)

Hence

$$\frac{dH_{F_1}(x)}{dx} = \frac{dt}{dx} = \frac{f_1[F_1^{-1}H_{F_1}(x)]}{1 - H_{F_1}(x)} \quad \text{and} \quad \frac{d}{dx}\int_0^{F_2^{-1}[H_{F_1}(x)]} \bar{F}_2(u)\,du = \frac{f_1[F_1^{-1}H_{F_1}(x)]}{f_2[F_2^{-1}\ H_{F_1}(x)]}.$$

But $F_1 \underset{c}{<} F_2$ implies

$$\frac{d}{dx}\ F_2^{-1}\ F_1(x) = \frac{f_1(x)}{f_2[F_2^{-1}F_1(x)]}$$

is nondecreasing in $0 \leqslant x \leqslant F_1^{-1}(1)$. Since $F_1^{-1}\ H_{F_1}(x)$ is nondecreasing in $0 \leqslant x \leqslant F_1^{-1}(1)$, a change of variable completes the argument.

Since continuous distributions can be approximated arbitrarily closely by absolutely continuous distributions, the proof of part (a) is complete.

To prove (b) we will need the following fundamental lemma.

FUNDAMENTAL LEMMA 2.3:

If $R(0) = 0, \dfrac{R(x)}{x}$ is nondecreasing in $x \geqslant 0$ and $0 \leqslant N(x) \leqslant \dfrac{1}{x} \int_0^x N(u)\,du$, then

$$\frac{\int_0^x N(u)\,dR(u)}{\int_0^x N(u)\,du} \tag{2.1}$$

is nondecreasing in $x \geqslant 0$. [Note that if $N(x)$ is nonincreasing, the assumption on $N(x)$ is automatically satisfied. For example, $N(x)$ could be the number surviving to time x.]

PROOF:

R can be approximated arbitrarily closely from below by positive linear combinations of simple functions of the form

$$R(x) = \begin{cases} 0 & x < x_o \\ x & x \geqslant x_0. \end{cases}$$

Hence we need only verify the lemma for simple functions. The general result follows from the Lebesgue monotone convergence theorem. For a simple function, R

$$\frac{\int_0^x N(u)\,dR(u)}{\int_0^x N(u)\,du} = \begin{cases} 0 & x < x_0 \\[2ex] \dfrac{x_0 N(x_0) + \int_{x_0}^x N(u)\,du}{\int_0^x N(u)\,du} & x \geqslant x_0. \end{cases}$$

Hence, for $x \geqslant x_0$

$$\frac{\int_0^x N(u)\,dR(u)}{\int_0^x N(u)\,du} = 1 + \frac{\left[x_0 N(x_0) - \int_0^{x_0} N(u)\,du\right]}{\int_0^x N(u)\,du}$$

By assumption, $N(x_0) \leqslant \dfrac{1}{x_0} \int_0^{x_0} N(u)\,du$ so that the lemma follows. | |

THEOREM 2.1, PART (B):

Let $R(x) = F_2^{-1} F_1(x)$. By assumption, $\dfrac{R(x)}{x}$ is nondecreasing in $0 \leqslant x \leqslant F_1^{-1}(1)$. Let $N(u) = \bar{F}_1(u)$, $x = F_1^{-1}(t)$ and substitute in (2.1) to obtain

$$\frac{\int_0^{F_1^{-1}(t)} \bar{F}_1(u)\,dR(u)}{\int_0^{F_1^{-1}(t)} \bar{F}_1(u)\,du}.$$

Let $v = F_2^{-1}\ F_1(u) = R(u)$ so that the numerator becomes $\int_0^{F_1^{-1}(t)} \bar{F}_1(u)\ dR(u) = \int_0^{F_2^{-1}(t)} \bar{F}_2(v)\,dv$. It follows from the Fundamental Lemma that

$$\frac{H_{F_2}^{-1}(t)}{H_{F_1}^{-1}(t)} = \frac{\int_0^{F_2^{-1}(t)} \bar{F}_2(v)\,dv}{\int_0^{F_1^{-1}(t)} \bar{F}_1(u)\,du}$$

is nondecreasing in $0 \leqslant t \leqslant 1$ or $\dfrac{H_{F_2}^{-1}H_{F_1}(x)}{x}$ is nondecreasing in $0 \leqslant x \leqslant F_1^{-1}(1)$, i.e.,

$$H_{F_1} \underset{*}{\leqslant} H_{F_2}.||$$

$\mathbf{H_{F_1} \underset{*}{\leqslant} H_{F_2}}$ DOES NOT IMPLY $\mathbf{F_1 \underset{*}{\leqslant} F_2}$

Let $G(x) = 1-e^{-x}$ so that $H_G(x) = x$ for $0 \leqslant x \leqslant 1$. It is easy to find examples such that $H_F \underset{*}{\leqslant} H_G$ but $F \underset{*}{\nless} G$; i.e., F is not IFRA. Note that for $0 < t_1 < 1$, $c = -log(1-t_1)-t_1 > 0$. Hence

$$\bar{F}(x) = \begin{cases} 1 & 0 \leqslant x < t_1 \\ e^{-(c+x)} & x \geqslant t_1 \end{cases}$$

is not IFRA since

$$\frac{R(x)}{x} = \begin{cases} 0 & 0 \leqslant x < t_1 \\ \dfrac{c}{x}+1 & x \geqslant t_1 \end{cases}$$

is decreasing for $x \geqslant t_1$.

But

$$H_F^{-1}(t) = \begin{cases} t_1 & 0 \leqslant t < t_1 \\ t & t_1 \leqslant t \leqslant 1, \end{cases}$$

is anti-starshaped; i.e., $\dfrac{H_F^{-1}(t)}{t}$ is nonincreasing in $0 < t < 1$ so that $H_F \underset{*}{<} H_G$.

The significance of this example is that an anti-starshaped total time on test plot is *not* necessarily evidence that F is IFRA.

3. A MEASURE OF IFRness

Figures 1 and 2 show scaled total time on test transformations for various parametric families of failure distributions. In each case a single shape parameter provides a measure of departure from exponentiality.

By Part (b) of Corollary 2.2, the area $\int_0^1 H_F^{-1}(u)\,du$ could also provide a measure of IFRness since if $F_1 \underset{c}{<} F_2$ and $\int_0^\infty x\,dF_1(x) = \int_0^\infty x\,dF_2(x)$, then

$$\int_0^1 H_{F_1}^{-1}(u)\,du \geqslant \int_0^1 H_{F_2}^{-1}(u)\,du\ .$$

If F has mean θ, then $\int_0^1 H_F^{-1}(u)\,du = \int_0^\theta x\,dH_F(x)$, so that the mean of H_F, the inverse of the transform of F provides a measure of the IFRness of F. The following lemma provides an easy means for calculating $\int_0^\theta x\,dH_F(x)$.

LEMMA:

If $\int_0^\infty x dF(x) < \infty$, then

$$\int_0^\infty x dH_F(x) = 2 \int_0^\infty x[1 - F(x)] dF(x) .$$

PROOF:

Since $\int_0^\infty x dH_F(x) = \int_0^1 H_F^{-1}(u) du$, we integrate the latter by parts of obtain

$$\int_0^1 H_F^{-1}(u) du = \int_0^1 F^{-1}(u) du - \int_0^1 t(1 - t) dF^{-1}(t) .$$

Integrate by parts again to obtain

$$-\int_0^1 t(1 - t) dF^{-1}(t) = \int_0^1 F^{-1}(u)(1 - 2u) du$$

so that

$$\int_0^1 H_F^{-1}(t) dt = 2 \int_0^1 (1 - u) F^{-1}(u) du = 2 \int_0^\infty x[1 - f(x)] dF(x)$$

by a change of variable.| |

EXAMPLES:

For $F(x) = 1 - e^{-(\lambda x)^\alpha}$ with mean, say θ,

$$\frac{1}{\theta} \int_0^\infty x dH_F(x) = 1/2^{1/\alpha} .$$

For the gamma distribution

$$F(x) = \int_0^x \frac{\lambda^k u^{k-1} e^{-\lambda u}}{(k - 1)!} du , \quad k = 1, 2, \ldots$$

with mean $\theta = \dfrac{k}{\lambda}$,

$$\frac{1}{\theta} \int_0^\infty x dH_F(x) = \sum_{i=0}^{k-1} \binom{i+k}{k} \frac{1}{2^{i+k}} .$$

The numerical relationship between $\frac{1}{\theta} \int_0^\infty x dH_F(x)$ and the shape parameter for Weibull and gamma distributions is shown in the following table. This is one way of relating Weibull and gamma distributions. The basic distinction between Weibull and gamma distributions lies in the behavior of their respective failure rates for large values of the time variable. Comparing Figures 1 and 2, we see that the gamma transform is nearly linear for $t \geqslant .50$ while the Weibull transform still exhibits curvature for $t \geqslant .50$.

TABLE 2 — *Relationship Between Measures of "IFRness"*

$\frac{1}{\theta}\int_0^\infty x dH_F(x)$	Weibull α	Gamma k
.20	.43	
.25	.50	
.30	.58	
.35	.66	
.40	.76	
.45	.87	
.50	1	1
.55	1.16	
.60	1.36	
.63	1.50	2
.65	1.61	
.69	1.87	3
.70	1.94	
.73	2.20	4
.75	2.41	5
.80	3.11	
.85	4.27	
.90	6.58	
.95	13.51	

ACKNOWLEDGEMENT

I would like to acknowledge helpful conversations with Frank Proschan and J. Sethuraman at Florida State University during the academic year 1975-76.

REFERENCES

[1] Barlow, R. E. and R. Campo, "Total time on Test Processes and Applications to Failure Data Analysis," in *Reliability and Fault Tree Analysis*, edited by R. E. Barlow, J. Fussell and N. Singpurwalla. Conference volume published by SIAM, Philadelphia (1975).

[2] Barlow, R. E., D. J. Bartholomew, J. M. Bremner and H. D. Brunk, *Statistical Inference Under Order Restrictions* (John Wiley and Sons, 1972).

[3] Birnbaum, Z. W., J. Esary and A. W. Marshall, "Stochastic Characterization of Wearout for Components and Systems," Annals of Mathematical Statistics, Vol. 37, pp. 816-825 (1966).

[4] Barlow, R. E. and F. Proschan, *Statistical Theory of Reliability and Life Testing* (Holt, Rinehart and Winston, 1975).

[5] Bergman, B., "Age Replacement and the TTT-Concept," Department of Mathematical Statistics, University of Lund, Lund, Sweden.

SIMULATION OF NONHOMOGENEOUS POISSON PROCESSES BY THINNING

P. A. W. Lewis*

Naval Postgraduate School
Monterey, California

G. S. Shedler

IBM Research Laboratory
San Jose, California

ABSTRACT

A simple and relatively efficient method for simulating one-dimensional and two-dimensional nonhomogeneous Poisson processes is presented. The method is applicable for any rate function and is based on controlled deletion of points in a Poisson process whose rate function dominates the given rate function. In its simplest implementation, the method obviates the need for numerical integration of the rate function, for ordering of points, and for generation of Poisson variates.

1. INTRODUCTION

The one-dimensional nonhomogeneous Poisson process (see e.g. [6], pp. 28-29; [4], pp. 94-101) has the characteristic properties that the numbers of points in any finite set of nonoverlapping intervals are mutually independent random variables, and that the number of points in any interval has a Poisson distribution. The most general nonhomogeneous Poisson process can be defined in terms of a monotone nondecreasing right-continuous function $\Lambda(x)$ which is bounded in any finite interval. Then the number of points in any finite interval, for example $(0, x_0]$, has a Poisson distribution with parameter $\mu_0 = \Lambda(x_0) - \Lambda(0)$. In this paper it is assumed that $\Lambda(x)$ is *continuous*, but *not* necessarily absolutely continuous. The right derivative $\lambda(x)$ of $\Lambda(x)$ is called the rate function of the process; $\Lambda(x)$ is called the integrated rate function and has the interpretation that for $x \geqslant 0$, $\Lambda(x) - \Lambda(0) = E[N(x)]$, where $N(x)$ is the total number of points in $(0, x]$. Note that $\lambda(x)$ may jump at points at which $\Lambda(x)$ is not absolutely continuous. In contrast to the homogeneous Poisson process, i.e., $\lambda(x)$ is a constant (usually denoted by λ), the intervals between the points in a one-dimensional nonhomogeneous Poisson process are neither independent nor identically distributed.

Applications of the one-dimensional nonhomogeneous Poisson process include modelling of the incidence of coal-mining disasters [6], the arrivals at an intensive care unit [12], transaction processing in a data base management system [14], occurrences of major freezes in Lake

*Support under Office of Naval Research Grant NR-42-343 is gratefully acknowledged.

Constance [20], and geomagnetic reversal data [19]. The statistical analysis of trends in a one-dimensional nonhomogeneous Poisson process, based on the assumption of an exponential polynomial rate function, is discussed by [6], [5], [12], and [14].

There are a number of methods for simulating the nonhomogeneous Poisson process which we review briefly.

(i) Time-scale transformation of a homogeneous (rate one) Poisson process via the inverse of the (continuous) integrated rate function $\Lambda(x)$ constitutes a general method for generation of the nonhomogeneous Poisson process (cf., [4], pp. 96-97). This method is based on the result that $X_1, X_2, \ldots$, are the points in a nonhomogeneous Poisson process with continuous integrated rate function $\Lambda(x)$ if and only if $X_1' = \Lambda(X_1)$, $X_2' = \Lambda(X_2), \ldots$, are the points in a homogeneous Poisson process of rate one. The time-scale transformation method is a direct analogue of the inverse probability integral transformation method for generating (continuous) nonuniform random numbers. For many rate functions, inversion of $\Lambda(x)$ is not simple and must be done numerically; cf., [7] and [17]. The resulting algorithm for generation of the nonhomogeneous Poisson process may be far less efficient than generation based on other methods; see e.g., [13], [15], and [17] for discussions of special methods for efficiently generating the nonhomogeneous Poisson process with log-linear and log-quadratic rate functions.

(ii) A second general method for generating a nonhomogeneous Poisson process with integrated rate function $\Lambda(x)$ is to generate the intervals between points individually, an approach which may seem more natural in the event-scheduling approach to simulation. Thus, given the points $X_1 = x_1, X_2 = x_2, \ldots, X_i = x_i$, with $X_1 < X_2 < \ldots < X_i$, the interval to the next point, $X_{i+1} - X_i$, is independent of $x_1, \ldots, x_{i-1}$ and has distribution function $F(x) = 1 - \exp[-\{\Lambda(x_i + x) - \Lambda(x_i)\}]$. It is possible to find the inverse distribution function $F^{-1}(\cdot)$, usually numerically, and generate $X_{i+1} - X_i$ according to $X_{i+1} - X_i = F^{-1}(U_i)$, where U_i is a uniform random number on the interval (0,1]. Note, however, that this not only involves computing the inverse distribution function for each interval $X_{i+1} - X_i$, but that each distribution has different parameters and possibly a different form. An additional complication is that $X_{i+1} - X_i$ is not necessarily a proper random variable, i.e., there may be positive probability that $X_{i+1} - X_i$ is infinite. It is necessary to take this into account for each interval $X_{i+1} - X_i$ before the inverse probability integral transformation is applied. The method is therefore very inefficient with respect to speed, more so than the time-scale transformation method.

(iii) In a third method, simulation of a nonhomogeneous Poisson process in a fixed interval $(0, x_0]$ can be reduced to the generation of a Poisson number of order statistics from a fixed density function by the following result (cf., [6], p. 45). If $X_1, X_2, \ldots, X_n$ are the points of the nonhomogeneous Poisson process in $(0,x_0]$, and if $N(x_0) = n$, then conditional on having observed $n(> 0)$ points in $(0, x_0]$, the X_i are distributed as the order statistics from a sample of size n from the distribution function $\{\Lambda(x) - \Lambda(0)\}/\{\Lambda(x_0) - \Lambda(0)\}$, defined for $0 < x \leqslant x_0$. Generation of the nonhomogeneous Poisson process based on order statistics is in general more efficient (with respect to speed) than either of the previous two methods. Of course, a price is paid for this greater efficiency. First, it is necessary to be able to generate Poisson variates, and second, more memory is needed than in the interval-by-interval method in order to store the sequence of points. Enough memory must be provided so that with very high probability the random numbers of points generated in the interval can be stored. Recall that the number of points in the interval $(0, x_0]$ has a Poisson distribution with mean

$\mu_0 = \Lambda(x_0) - \Lambda(0)$. Memory of size, e.g., $\mu_0 + 4\mu_0^{1/2}$ wll ensure that overflow will occur on the average in only one out of approximately every 40,000 realizations. This probability is small enough so that in the case of overflow, the realization of the process can generally be discarded.

(iv) Again, there is a very particular and very efficient method for simulation of nonhomogeneous Poisson processes with log-linear rate function [13] which, at the cost of programming complexity and memory, can be used to obtain an efficient simulation method for other rate functions, as in [15].

In this paper a new method is given for simulating a nonhomogeneous Poisson process which is not only conceptually simple, but is also computationally simple and relatively efficient. In fact, at the cost of some efficiency, the method can be applied to simulate the given nonhomogeneous Poisson process *without the need for numerical integration or routines for generating Poisson variates.* Used in conjunction with the special methods given in [13] and [15], the method can be used to generate quite efficiently nonhomogeneous Poisson processes with rather complex rate functions, in particular combinations of long-term trends and fixed-cycle effects. The method is also easily extended to the problem of generating the two-dimensional nonhomogeneous Poisson process.

2. SIMULATION OF ONE-DIMENSIONAL NONHOMOGENEOUS POISSON PROCESSES

Simulation of a nonhomogeneous Poisson process with general rate function $\lambda(x)$ in a fixed interval can be based on thinning of a nonhomogeneous Poisson process with rate function $\lambda^*(x) \geqslant \lambda(x)$. The basic result is

THEOREM 1: Consider a one-dimensional nonhomogeneous Poisson process $\{N^*(x):x \geqslant 0\}$ with rate function $\lambda^*(x)$, so that the number of points, $N^*(x_0)$, in a fixed interval $(0, x_0]$ has a Poisson distribution with parameter $\mu_0^* = \Lambda^*(x_0) - \Lambda^*(0)$. Let $X_1^*, X_2^*, \ldots, X_{N^*(x_0)}^*$ be the points of the process in the interval $(0,x_0]$. Suppose that for $0 \leqslant x \leqslant x_0$, $\lambda(x) \leqslant \lambda^*(x)$. For $i = 1, 2, \ldots, n$, delete the point X_i^* with probability $1 - \lambda(X_i^*)/\lambda^*(X_i^*)$; then the remaining points form a nonhomogeneous Poisson process $\{N(x):x \geqslant 0\}$ with rate function $\lambda(x)$ in the interval $(0,x_0]$.

PROOF: Since $\{N^*(x):x \geqslant 0\}$ is a nonhomogeneous Poisson process and points are deleted independently, it is clear that the number of points in $\{N(x):x \geqslant 0\}$ in any set of nonoverlapping intervals are mutally independent random variables. Thus, it is sufficient to show that the number of points $N(a,b)$ in $\{N(x):x \geqslant 0\}$ in an arbitrary interval $(a,b]$ with $0 \leqslant a < b \leqslant x_0$ has a Poisson distribution with parameter $\Lambda(b) - \Lambda(a)$. Observe that with $p(a,b) = \{\Lambda(b) - \Lambda(a)\}/\{\Lambda^*(b) - \Lambda^*(a)\}$, we have the conditional probability:

$$(1) \qquad P\{N(a,b) = n \mid N^*(a,b) = k\} = \begin{cases} 1 & \text{if } n = k = 0 \\ \binom{k}{n} \{p(a,b)\}^n \{1 - p(a,b)\}^{k-n} & \text{if } k \geqslant n \geqslant 0 \\ & \text{and } k \geqslant 1 \\ 0 & \text{if } n \geqslant 1 \\ & \text{and } k < n \end{cases}$$

Equation (1) is a consequence of the well-known result that, conditional on k (>0) points in the interval $(a,b]$, the joint density of the k points in the process $\{N^*(x):x \geqslant 0\}$ is $\lambda^*(x_1) \ldots \lambda^*(x_k)/\{\Lambda^*(b) - \Lambda^*(a)\}^k$. The desired result is obtained in a straightforward manner from Equation (1) by removing the condition.

Theorem 1 is the basis for the method of simulating nonhomogeneous Poisson processes given in this paper.

ALGORITHM 1: One-dimensional nonhomogeneous Poisson process.

1. Generate points in the nonhomogeneous Poisson process $\{N^*:(x) \geqslant 0\}$ with rate function $\lambda^*(x)$ in the fixed interval $(0,x_0]$. If the number of points generated, n^*, is such that $n^* = 0$, exit; there are no points in the process $\{N(x):x \geqslant 0\}$.

2. Denote the (ordered) points by $X_1^*, X_2^*, \ldots, X_{n^*}^*$. Set $i = 1$ and $k = 0$.

3. Generate U_i, uniformly distributed between 0 and 1. If $U_i \leqslant \lambda(X_i^*)/\lambda^*(X_i^*)$, set k equal to $k+1$ and $X_k = X_i^*$.

4. Set i equal to $i+1$. If $i \leqslant n^*$, go to 3.

5. Return $X_1, X_2, \ldots, X_n$, where $n = k$, and also n.

(i) In the case where $\{N^*(x):x \geqslant 0\}$ is a homogeneous Poisson process with $\lambda^*(x) = \lambda^*$;

(ii) the minimum of $\lambda(x)$, say $\underline{\lambda}$, is known, and

(iii) generation of uniformly distributed variates is computationally costly,

considerable speedup can be obtained by noting that X_i^* is always accepted if $U_i \leqslant \underline{\lambda}/\lambda^*$. This obviates, in some cases, computation of $\lambda(x)$, which is the main source of inefficiency in the algorithm. Moreover, in this case $\lambda^* U_i/\underline{\lambda}$ can be used as the next uniformly distributed variate.

The method of thinning in this simple form, i.e., $\lambda^*(x) = \lambda^* = \max_{0 \leqslant x \leqslant x_0} \lambda(x)$, can also be used to provide an algorithm for generating a nonhomogeneous Poisson process on an interval-by-interval basis, as discussed in subsection (ii) of Section 1. The interval to the next point $X_{i+1} - X_i$ is obtained by generating and cumulating exponential (λ^*) random numbers $E_1^*, E_2^*, \ldots$, until for the first time $U_j \leqslant \lambda(X_i + E_1^* + \ldots E_j^*)/\lambda^*$, where the U_j are independent uniform random numbers between 0 and 1. This algorithm is considerably simpler than the interval-by-interval algorithm of Section 1 since it requires no numerical integration, only the availability of uniform random numbers.

3. DISCUSSION OF THE METHOD OF THINNING

(i) Relationship to acceptance-rejection method

The method of thinning of Algorithm 1 is essentially the obverse of the conditional method of Section 1, using conditioning and acceptance-rejection techniques to generate the random variables with density function $\lambda(x)/\{\Lambda(x) - \Lambda(0)\}$ (Lewis and Shedler, [15], Algorithm 3). The differences are subtle, but computationally important. In the acceptance-rejection method, it is first necessary to generate a Poisson variate with mean $\mu_0 = \Lambda(x_0) - \Lambda(0)$, and this involves an integration of the rate function $\lambda(x)$. Then the Poisson (μ_0) number, n, of variates generated by acceptance-rejection must be ordered to give $X_1, X_2, \ldots, X_n$.

(ii) Simplest form of the thinning algorithm

In the simplest form of the method of thinning, $\lambda^*(x)$ is taken to be a constant λ^*, so that, for instance, the points $X_1^*, X_2^*, \ldots, X_{n^*}^*$ can be generated by cumulating exponential (λ^*) variates until the sum is greater than x_0 (cf., [13], Algorithm 1). Thinning is then applied to the generated points. *No ordering, no integration of $\lambda(x)$ and no generator of Poisson variates is required.* Of course for both algorithms to be efficient, computation of $\lambda(x)$ and $\lambda^*(x)$ must be easy relative to computation of the inverse of $\Lambda(x)$.

(iii) Efficiency

For the thinning algorithm (as well as the algorithm based on conditioning and acceptance-rejection) efficiency, as measured by the number of points deleted, is proportional to $\mu_0/\mu_0^* = \{\Lambda(x_0) - \Lambda(0)\}/\{\Lambda^*(x_0) - \Lambda^*(0)\}$; this is the ratio of the areas between 0 and x_0 under $\lambda(x)$ and $\lambda^*(x)$. Thus, $\lambda^*(x)$ should be as close as possible to $\lambda(x)$ consistent with ease of generating the nonhomogeneous Poisson process $\{N^*(x):x \geqslant 0\}$.

(iv) An example: fixed cycle plus trend

To illustrate the applicability of the thinning algorithm, consider its use in conjunction with the algorithms given by [13] and [15] for log-linear and log-quadratic rate functions. Assume that it is necessary to simulate a nonhomogeneous Poisson process whose rate function increases quadratically with time but also has a fixed-period cycle, e.g.,

$$\lambda(x) = \exp\{\alpha_0 + \alpha_1 x + \alpha_2 x^2 + K \sin(\omega_0 x + \theta)\},$$

$$0 \leqslant x \leqslant x_0;\ K \geqslant 0;\ 0 < \theta \leqslant 2\pi;\ \omega_0 > 0.$$

This is the model found by Lewis [12] for arrivals at an intensive care unit, where there is a strong time-of-day effect. Thus if $\omega_0 = 2\pi/T_0$, then the period $T_0 = 1$ day. Computation of $\Lambda^{-1}(\cdot)$ is difficult. To determine $\lambda^*(x)$, note that

$$\lambda(x) \leqslant \lambda^*(x) = \exp\{\alpha_0 + K + \alpha_1 x + \alpha_2 x^2\},$$

and therefore

$$\lambda(x)/\lambda^*(x) = \exp[K\{1 - \sin(\omega_0 x + \theta)\}].$$

Thus in step 3 of Algorithm 1, U_i is compared to $\exp[K\{1 - \sin(\omega_0 X_i^* + \theta)\}]$. Equivalently, if unit exponential variates E_i are available, it is faster to compare E_i to $K\{1 - \sin(\omega_0 X_i^* + \theta)\}$, accepting X_i^* if $E_i > K\{1 - \sin(\omega_0 X_i^* + \theta)\}$.

The main computational expense here is generation of the E_i and computation of the sine function, both n^* times. The expense involved in computation of the sine function can be reduced by noting that the point X_i^* is always accepted if E_i is greater than $2K$. This will be a great saving if the cyclic effect is minor (K small). The number of E_i generated can be reduced by noting that if, in one step of the algorithm, E_i is observed to be greater than δ, then $E_i^* = E_i - \delta$ can be used as an (independent) unit exponential variate in the next step. The above procedure can be extended to the case of a trend with two fixed-period cycles, e.g., a time-of-day and a time-of-week effect.

4. SIMULATION OF TWO-DIMENSIONAL HOMOGENEOUS POISSON PROCESSES

The two-dimensional homogeneous Poisson process (of rate $\lambda > 0$) is defined by the properties that the numbers of points in any finite set of nonoverlapping regions having areas in the usual geometric sense are mutually independent, and that the number of points in any region of area A has a Poisson distribution with mean λA; see, e.g. [11], pp. 31-32. Note that the number of points in a region R depends on its area, but not on its shape or location. The homogeneous Poisson process arises as a limiting two-dimensional point process with respect to a number of limiting operations; cf., [8] and [9]. Properties of the process are given by [16]. Applications of the two-dimensional homogeneous Poisson process to problems in ecology and forestry have been discussed by Thompson [21] and Holgate [10]. The model also arises in connection with naval search and detection problems.

In considering the two-dimensional homogeneous Poisson process, projection properties of the process depend quite critically on the geometry of the regions considered. These projection properties are simple for rectangular and circular regions, and make simulation of the homogeneous process quite easy. We consider these two cases separately.

(i) Homogeneous Poisson processes in a rectangle

The following two theorems form the basis for simulation of the two-dimensional homogeneous Poisson process in a rectangle.

THEOREM 2: Consider a two-dimensional homogeneous Poisson process of rate λ, so that the number of points in a fixed rectangle $R = \{(x,y): 0 < x \leqslant x_0, 0 < y \leqslant y_0\}$ has a Poisson distribution with parameter $\lambda x_0 y_0$. If $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ denote the position of the points of the process in R, labelled so that $X_1 < X_2 < \ldots < X_N$, then $X_1, X_2, \ldots, X_N$ form a one-dimensional homogeneous Poisson process on $0 < x \leqslant x_0$ of rate λy_0. If the points are relabelled $(X_1', Y_1'), (X_2', Y_2'), \ldots, (X_N', Y_N')$ so that $Y_1' < Y_2' < \ldots < Y_N'$, then $Y_1', Y_2', \ldots, Y_N'$ form a one-dimensional homogeneous Poisson process on $0 < y \leqslant y_0$ of rate λx_0.

PROOF: The number of points in an interval on the x-axis, say $(a,b]$ is the number of points in the rectangle bounded by the lines $x = a$, $x = b$, $y = 0$, and $y = y_0$. This number is therefore independent of the number of points in any similar nonoverlapping rectangle bounded on the x-axis by $x = a'$, $x = b'$, i.e., the number of points in the interval $(a',b']$. This establishes the independent increment property for a one-dimensional Poisson process. The Poisson distribution of the number of points in $(a,b]$ follows from the fact that it is equal to the number of points in the rectangle bounded by $x = a$, $x = .b$, $y = 0$, and $y = y_0$, and the latter has a Poisson distribution wih parameter $\lambda y_0(b-a)$. An analogous argument shows that the process formed on the y-axis by $Y_1', Y_2', \ldots, Y_N'$ is Poisson.

Conditional properties of the Poisson process in a rectangle are established next. The important thing to note is that while the processes obtained by projection of the points onto the x and y axes are not independent, there is a type of conditional independence which makes it easy to simulate the two-dimensional process.

THEOREM 3: Assume that a two-dimensional homogeneous Poisson process of rate λ is observed in a fixed rectangle $R = \{(x,y): 0 < x \leqslant x_0, 0 < y \leqslant y_0\}$, so that the number of points in R, $N(R)$, has a Poisson distribution with parameter $\lambda x_0 y_0$. If $N(R) = n > 0$ and if $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ denote the points, labelled so that $X_1 < X_2 < \ldots < X_n$, then conditional on having observed n points in R, the $X_1, X_2, \ldots, X_n$ are uniform order statistics on $0 < x \leqslant x_0$, and $Y_1, Y_2, \ldots, Y_n$ are independent and uniformly distributed on $0 < y \leqslant y_0$, independent of the X_i.

PROOF: If there are N points in the rectangle, form N vertical strips, from 0 to y_0 and from X_i to $X_i + dx_i$, such that each strip contains only one of the N points. The position of Y_i on the vertical line through X_i is that of an event in a Poisson process of rate λdx_i, given that only one event occurs. But this means that Y_i is uniformly distributed between 0 and y_0. Moreover, this is true irrespective of where X_i occurs; therefore Y_i is independent of X_i. Also, occurrences in all N strips are independent, and therefore Y_i is independent of the other Y_j and X_j positions $j \neq i$. Thus, the Y_i are a random sample of size N from a uniform $(0,y_0)$ distribution, independent of the X_i. Now condition on $N = n$ (> 0); since by Theorem 2 the X_i form a Poisson process they are, by well-known results, order statistics from a uniform $(0,x_0)$ sample and are independent of the fixed size Y_i population; thus the pairs (X_i, Y_i) are mutually independent.

COROLLARY: Denote the Poisson points by $(X_1, Y_1), (X_2, Y_2), \ldots$, where the index does not necessarily denote an ordering on either axis. Conditionally, the pairs $(X_1, Y_1), \ldots, (X_N, Y_N)$ are independent random variables. Furthermore, for each pair, X_i is distributed uniformly between 0 and x_0, independently of Y_i, which is uniformly distributed between 0 and y_0.

From the two theorems, the following simulation procedure is obtained.

ALGORITHM 2: Two-dimensional homogeneous Poisson process in a rectangle.

1. Generate points in the one-dimensional homogeneous Poisson process of rate λy_0 on $(0, x_0]$. If the number of points generated, n, is such that $n = 0$, exit; there are no points in the rectangle.

2. Denote the points generated by $X_1 < X_2 < \ldots < X_n$.

3. Generate $Y_1, Y_2, \ldots, Y_n$ as independent, uniformly distributed random numbers on $(0,y_0]$

4. Return $(X_1,Y_1), (X_2,Y_2), \ldots, (X_n,Y_n)$ as the coordinates of the two-dimensional homogeneous Poisson process in the rectangle, and n.

Note that generation of the points $X_1, X_2, \ldots, X_N$ in Steps 1 and 2 can be accomplished by cumulating exponential (λy_0) random numbers. Alternatively, after generating a Poisson random number $N=n$ (with parameter $\lambda x_0 y_0$), n independent, uniformly distributed random numbers on $(0,x_0]$ can be ordered; see [13], p. 502.

Another algorithm for generation of the two-dimensional Poisson process in a rectangle can be based on the Corollary to Theorem 3.

(ii) Homogeneous Poisson processes in a circle

The following two theorems form the basis for simulation of the two-dimensional homogeneous Poisson process in a fixed circle of radius r_0.

Fix the origin and initial line of polar coordinates r and θ so that the origin is the center of the circle and the initial line is horizontal. We consider the projection of the points (R_i, θ_i), of the Poisson process circularly onto the r-axis (R_i) and radially onto the circumferential θ-axis (θ_i). The number of points projected onto the r-axis in the interval $(0,r]$, where $r \leqslant r_0$, is the number of points in the circle of radius r and area πr^2; thus the number of points in $(0,r]$ has a Poisson distribution with parameter $\lambda \pi r^2$. Consequently, if the projection process on the r-axis is a Poisson process, it must have integrated rate function $\Lambda(r) = \lambda \pi r^2$, with $\Lambda(0) = 0$.

Similarly, the number of points on the circumferential arc of the fixed circle (radius r_0) from 0 to θ is the number of points in the sector of the circle defined by radial lines at angles 0 and θ; thus the number of points on the arc from 0 to θ has a Poisson distribution with parameter $\lambda \pi r_0^2 \times \dfrac{\theta}{2\pi} = \theta \lambda r_0^2/2$. Accordingly, if the projection process on the θ-axis is a Poisson process, it must have integrated rate function $\Lambda(\theta) = \theta \lambda r_0^2/2$, with $\Lambda(0) = 0$.

We now assert that the projection processes are in fact Poisson processes. Since proofs of these theorems are directly analogous to the proofs of Theorems 2 and 3, they are omitted.

THEOREM 4: Consider a two-dimensional homogeneous Poisson process of rate λ so that the number N of points in a fixed circular area C of radius r_0 and area πr_0^2 has a Poisson distribution with parameter $\lambda \pi r_0^2$. If $(R_1,\theta_1), (R_2,\theta_2), \ldots, (R_N,\theta_N)$ denote the points of the process in C, labelled so that $R_1 < R_2 < \ldots < R_N$, then $R_1, R_2, \ldots, R_N$ form a one-dimensional nonhomogeneous Poisson process on $0 \leqslant r \leqslant r_0$ with rate function $\lambda(r) = 2\pi\lambda r$. If the points are relabelled $(R_1',\theta_1'), (R_2',\theta_2'), \ldots, (R_N',\theta_N')$ so that $\theta_1' < \theta_2' < \ldots < \theta_N'$, then $\theta_1', \theta_2', \ldots, \theta_N'$ form a one-dimensional homogeneous Poisson process on $0 < \theta \leqslant 2\pi$ of rate $\lambda r_0^2/2$.

THEOREM 5: Assume that a two-dimensional Poisson process of rate λ is observed in a fixed circular area C of radius r_0 so that the number of points in C, $N(C)$, has a Poisson distribution with parameter $\lambda \pi r_0^2$. If $N(C) = n > 0$ and if $(R_1,\theta_1), (R_2,\theta_2), \ldots, (R_n,\theta_n)$ with $R_1 < R_2 < \ldots < R_n$ denote the points, then conditional on having observed n points in C, the $R_1, R_2, \ldots, R_n$ are order statistics from the density $f(r) = 2r/r_0^2$ concentrated on $0 \leqslant r \leqslant r_0$, and $\theta_1, \theta_2, \ldots, \theta_n$ are independent and uniformly distributed on $0 < \theta \leqslant 2\pi$, independent of the R_i. These theorems lead to the following simulation procedure.

ALGORITHM 3: Two-dimensonal homogeneous Poisson process in a circular area.

1. Generate n as a Poisson random number with parameter $\lambda \pi r_0^2$. If $n = 0$, exit; there are no points in C.

2. Generate n independent random numbers having density function $f(r) = 2r/r_0^2$ and order to obtain $R_1 < R_2 < \ldots < R_n$.

3. Generate $\theta_1, \theta_2, \ldots, \theta_n$ independent, uniformly distributed random numbers on $(0,2\pi]$.

4. Return $(R_1,\theta_1), (R_2,\theta_2), \ldots, (R_n,\theta_n)$, and n.

Note that the wedge-shaped density $2r/r_0^2$ can be generated by scaling the maximum of two independent uniform (0,1) random numbers.

Direct generation of homogeneous Poisson points in non-circular or non-rectangular regions is difficult. The processes obtained by projection of the points on the two axes are nonhomogeneous Poisson processes with complex rate functions determined by the geometry of the region. However, the conditional independence which is found in circular and rectangular regions (Theorems 3 and 5) for the processes on the two axes is not present. In particular, given that there are n points $(X_1, Y_1), \ldots, (X_n, Y_n)$ in a non-rectangular region, the pairs (X_i, Y_i) are mutually independent, but X_i is in general not independent of Y_i, $i = 1, \ldots, n$. Therefore, it is simpler to enclose the region in either a circle or a rectangle, generate a homogeneous Poisson process in the enlarged area, and subsequently exclude points outside of the given region.

5. SIMULATION OF TWO-DIMENSIONAL NONHOMOGENEOUS POISSON PROCESSES

The two-dimensional nonhomogeneous Poisson process $\{N(x,y): x \geqslant 0,\ y \geqslant 0\}$ is specified by a positive rate function $\lambda(x,y)$ which, for simplicity, is assumed here to be continuous. Then the process has the characteristic properties that the numbers of points in any finite set of nonoverlapping regions having areas in the usual geometric sense are mutually independent, and that the number of points in any such region R has a Poisson distribution with mean $\Lambda(R)$; here $\Lambda(R)$ denotes the integral of $\lambda(x,y)$ over R, i.e., over the entire area of R.

Applications of the two-dimensional nonhomogeneous Poisson process include problems in forestry and naval search and detection. The use of the process as a model for the pattern of access to the storage subsystem of a computer system will be reported elsewhere. Detection and statistical analysis of trends in the two-dimensional nonhomogeneous Poisson process is discussed by Rantschler [18].

Theorem 1 dealing with thinning of one-dimensional nonhomogeneous Poisson processes generalizes to two-dimensional nonhomogeneous Poisson processes. Thus, suppose that $\lambda(x,y) \leqslant \lambda^*(x,y)$ in a fixed rectangular region of the plane. If a nonhomogeneous Poisson process with rate function $\lambda^*(x,y)$ is thinned according to $\lambda(x,y)/\lambda^*(x,y)$ (i.e., each point (X_i, Y_i) is deleted independently if a uniform (0,1) random number U_i is greater than $\lambda(X_i, Y_i)/\lambda^*(X_i, Y_i)$), the result is a nonhomogeneous Poisson process with rate function $\lambda(x,y)$. The proof is a direct analogue of the proof for the one-dimensional case.

The nonhomogeneous Poisson process with rate function $\lambda(x,y)$ in an arbitrary but fixed region R can be generated by enclosing the region R either in a rectangle or a circle, and applying Algorithm 2 or Algorithm 3. The following procedure assumes that the region R has been enclosed in a rectangle R^*, and that $\lambda^* = \max\{\lambda(x,y): x,y \in R\}$ has been determined; here the bounding process is homogeneous with rate λ^* in the rectangle R^*.

ALGORITHM 4: Two-dimensional nonhomogeneous Poisson process.

1. Using Algorithm 2, generate points in the homogeneous Poisson process of rate λ^* in the rectangle R^*. If the number of points, n^*, is such that $n^* = 0$, exit; there are no points in the nonhomogeneous Poisson process.

2. From the n^* points generated in 1, delete the points that are not in R, and denote the remaining points by $(X_1^*, Y_1^*), (X_2^*, Y_2^*), \ldots, (X_m^*, Y_m^*)$ with $X_1^* < X_2^* < \ldots < X_m^*$. Set $i = 1$ and $k = 0$.

3. Generate U_i uniformly distributed between 0 and 1. If $U_i \leqslant \lambda(X_i^*, Y_i^*)/\lambda^*$, set $k = k+1$, $X_k = X_i^*$ and $Y_k = Y_i^*$.

4. Set i equal to $i+1$. If $i \leqslant m^*$, go to 3.

5. Return (X_1, Y_1), (X_2, Y_2), $\ldots$, (X_n, Y_n), where $n = k$, and n.

It is not necessary that the bounding process have a constant rate λ^*. Theorems 2 and 4 can be generalized to certain cases where the process is nonhomogeneous (cf., [3]), for instance $\lambda(x,y) = \rho(x)\,\psi(y)$. Thus, a tighter bounding process which is nonhomogeneous may possibly be obtained. It is not simple to see how much efficiency could be gained by doing this, as opposed to using a two-dimensional homogeneous Poisson process for the bounding process. Again, as in the one-dimensional case, savings in computing $\lambda(x,y)$ can be obtained by computing its minimum beforehand, and the U_i's can be reused by scaling.

6. COMPARISONS AND CONCLUDING REMARKS

The method of thinning presented in this paper for simulating one-dimensional and two-dimensional nonhomogeneous Poisson processes with given rate function can be carried out in a computationally simple way by using a bounding process which is homogeneous with a rate function equal to the maximum value of the given rate funciton. No numerical integration, ordering, or generation of Poisson variates is required, only the ability to evaluate the given rate function. The thinning algorithm appears to be particularly attractive in the two-dimensional case where there seem to be no competing algorithms.

The thinning algorithm can also be implemented more efficiently at the cost of programming complexity and by using a nonhomogeneous bounding process. In particular the method can be used in conjunction with the special algorithms given by [13] and [15].

It is also possible to extend the method of thinning to simulation of doubly stochastic or conditioned Poisson processes. This will be discussed elsewhere.

REFERENCES

[1] Ahrens, J. and U. Dieter, "Computer Methods for Sampling from Gamma, Beta, Poisson and Binomial Distributions," Computing 12, 223-46 (1974).

[2] Ahrens, J. and U. Dieter, *Non Uniform Random Numbers*, (Technische Hochschule in Graz, Graz, Austria, 1973).

[3] Bartlett, M. S. "The Spectral Analysis of Two-dimensional Point Processes," Biometrika 51, 299-311 (1964).

[4] Cinlar, E., *Introduction to Stochastic Processes*, (Prentice Hall, Englewood Cliffs, New Jersey, 1975).

[5] Cox, D. R., "The Statistical Analysis of Dependencies in Point Processes," in *Stochastic Point Processes*, Ed. P. A. W. Lewis, pp. 55-56 (Wiley, New York, 1972).

[6] Cox, D. R. and P. A. W. Lewis, *The Statistical Analysis of Series of Events*, Methuen, London (1966).

[7] Gilchrist, R., "Some Problems Encountered in Generating Realisations from a Useful Non-Homogeneous Poisson Process," in *Proceedings of the European Congress of Statisticians*, Ed. J. R. Barra et al. (Amsterdam, North Holland, 1976).

[8] Goldman, J. R., "Stochastic Point Processes: Limit Theorems," Annals of Mathematical Statistics 38, 771-79 (1967).

[9] Goldman, J. R., "Infinitely Divisible Point Processes in R^n," Journal of Mathematical Analysis and Applications *17*, 133-46 (1967).

[10] Holgate, P., "The Use of Distance Methods for the Analysis of Spatial Distibutions of Points," in *Stochastic Point Processes*, Ed. P. A. W. Lewis, pp. 122-25 (Wiley, New York, 1972).

[11] Karlin, S. and H. M. Taylor, *A First Course in Stochastic Processes* (Academic Press, New York, 1975).

[12] Lewis, P. A. W., "Recent Advances in the Statistical Analysis of Univariate Point Processes," in *Stochastic Point Processes*, Ed. P. A. W. Lewis, pp. 1-54 (Wiley, New York, 1972).

[13] Lewis, P. A. W. and G. S. Shedler, "Simulation of Non-Homogeenous Processes with Log-Linear Rate Function." Biometrika 63, 501-05 (1976).

[14] Lewis, P. A. W. and G. S. Shedler, "Statistical Analysis of Non-Stationary Series of Events in a Data Base System," IBM Journal of Research and Development 20, 465-82 (1976).

[15] Lewis, P. A. W. and G. S. Shedler, "Simulation of Nonhomogeneous Poisson Processes with Degree-Two Exponential Polynomial Rate Function," IBM Research Report RJ 1953, San Jose, California (1977). To appear in Operations Research.

[16] Miles, R. E., "On the Homogeneous Planar Poisson Point Process," Mathematical Biosciences 6, 85-127 (1970).

[17] Patrow, M. L., "A Comparison of Two Algorithms for the Simulation of Non-Homogeneous Poisson Processes with Degree-Two Exponential Polynomial Intensity Function," M. S. Thesis, Dept. of Operations Research, Naval Postgraduate School, Monterey, California (National Technical Informaiton Service, Springfield, Virgina AD-A047164) (1977).

[18] Rantschler, R. D., "The Detection and Estimation of Trends in Two-Dimensional Poisson Processes," M.S. Thesis, Dept. of Operations Research, Naval Postgraduate School, Monterey, California (National Technical Information Service, Springfield, Virginia; AD-A76136) (1973).

[19] Reyment, R. A., "Geomagnetic Reversal Data Statistically Appraised," Tectonophysics 31, 73-91 (1976).

[20] Steinijans, V., "A Stochastic Point Process Model of the Occurrence of Major Freezes in Lake Constance," Applied Statistics 25, 58-61 (1976).

[21] Thompson, H. R., "Spatial Point Processes with Applications to Ecology," Biometrika 42, 102-15 (1955).

[22] Vere-Jones, D., "Stochastic Models for Earthquake Occurrence, Journal of Royal Statistical Society, B 32, 1-62 (1970).

BOUNDS ON OPTIMAL COST FOR A REPLACEMENT PROBLEM WITH PARTIAL OBSERVATIONS*

Chelsea C. White, III

Department of Applied Mathematics and Computer Science
University of Virginia
Charlottesville, Virginia

ABSTRACT

This paper characterizes the structure of optimal strategies for a replacement problem for two special cases of observation quality. It is shown that when the state of the system is either completely observed or completely unobserved at every decision epoch by the controller, reasonable assumptions are sufficient for the existence of optimal replacement strategies composed of policies having a generalized, control-limit form. These structural results are of particular interest since the optimal cost functions for the two special cases represent bounds on the optimal cost function for the general partially observed case, significant computational simplification can result for the two special cases due to their optimal strategy structure, and optimal strategies possessing a control-limit structure do not necessarily exist for the general partially observed case.

1. INTRODUCTION

In this paper the structural properties of the optimal replacement strategy for two special cases of a general replacement problem are characterized. The system is modeled by a Markov process. The controller can either replace or do nothing at each decision epoch. The system is assumed partially observed; that is, only the realization of a stochastic process, probabilistically related to the system's state process, is made available to the controller just prior to decision epochs. Thus, the replacement problem is modeled by the partially observed Markov decision process [1,7,8].

The two special cases considered are the completely observed case, where the controller knows the system's state exactly, and the completely unobserved case, where the controller receives no information regarding the state process through the observation process. For both cases it is shown that a generalized notion of the increasing failure rate [3] and reasonable assumptions on the cost structure are sufficient for the existence of optimal strategies having a generalized, control-limit structure.

Interest in these special cases is two-fold. First, they represent valuable models of system replacement. Second, the optimal expected cost functions for the completely observed and the completely unobserved cases represent lower and upper bounds, respectively, on the optimal expected cost function for the general partially observed problem [1]. This second result is

*This work has been supported by NSF Grant ENG-75-19692.

particularly interesting in light of the following facts. First, the optimal cost and strategy for the general partially observed problem are not guaranteed to have a monotone structure, as has been shown in [11] by counterexample. Second, the computational requirements for the general case can be substantial [6,8]. Third, considerable computational reduction may be induced by the optimal cost and strategy structure for the two cases of interest.

Section 2 presents the general, partially observed, Markov decision model of system replacement. The finite-horizon, discounted cost, and average cost criteria are considered. Applying results due to Porteus [4], the existence of optimal strategies having control-limit structures is proved in Section 3 for the two special cases. Emphasis is placed on the less well-known, completely unobserved problem. Computational aspects of these two cases are briefly discussed.

2. PROBLEM FORMULATION AND PRELIMINARIES FOR THE GENERAL CASE

Let the stochastic process $\{s(t),\ t = 0,1, \ldots\}$, having finite state space S, model a system subject to Markov deterioration. Let $\leqslant$ be a partial order on S, describing the relative deterioration of the system, i.e. if $i \leqslant i'$ then i' is a less desirable system state than i. We assume that S is only partially ordered to allow $\{s(t),\ t = 0,1, \ldots\}$ to be a vector process which could serve as a model of a multi-component system or a system subject to semi-Markov deterioration. Let 0 be a minimal element in S; i.e., $0 \leqslant i$ for all $i \in S$. The state 0 therefore corresponds to the "best" or "newly replaced" state. Assume the a priori probabilities $p^0 = \{p_i^0\} \in \Omega = \left\{x : x_i \geqslant 0, \sum_{i \in S} x_i = 1\right\}$ are given, where $p_i^0 = P[s(0) = i]$.

The controller observes the state of the system at each time $t = 1, 2, \ldots$. The observation may be noise-corrupted and is assumed to be the realization of a random variable $z(t)$ at time t. The resulting process $\{z(t),\ t = 1,2, \ldots\}$ is called the observation process, which is assumed to have a finite state space.

A decision is made at each time $t = 0, 1, \ldots, n-1$ to replace the system or to keep it until the next decision epoch, where $n \leqslant \infty$ represents the terminal time of the problem. The decision to replace is equivalent to guaranteeing that the state of the system at the next decision epoch will be 0. Let $u(t) \in D = \{0,1\}$ represent the decision made at time t, where 0=do nothing and 1=replace. It is assumed that the controller bases all decisions on all former decisions made and all past and present observations of the system; hence,

$$u(t) = \gamma_t[u(0), \ldots, u(t-1), z(1), \ldots, z(t)]. \tag{1}$$

The processes $\{s(t),\ t = 0,1, \ldots\}$, $\{z(t),\ t = 1,2, \ldots\}$, and $\{u(t),\ t = 0,1, \ldots\}$ are all related by the conditional probabilities $p_{ij}\ (y) = P[s(t+1) = j \mid s(t) = i,\ u(t) = y]$ and $q_{jk}(y) = P[z(t+1) = k \mid s(t+1) = j,\ u(t) = y]$, where $p_{ij} = p_{ij}(0)$ and $p_{ij}(1) = 1\ j = 0$.

Let $r[s(t),\ u(t)]$ be the cost accrued at time $0 \leqslant t < n$. Additionally, when the problem horizon is finite, assume a terminal cost $r_0[s(n)]$ is accrued at the terminal time. All costs are assumed to be nonnegative.

The finite horizon problem is to select a sequence of policies γ_t, $t = 0, 1, \ldots n-1$, satisfying (1) which minimizes the criterion $E\{\Sigma_{t=0}^{n-1}\ \beta^t\ r[s(t),\ u(t)] + \ \beta^n r_0[s(n)] \mid p^0\}$, where

$\beta \in [0,1]$ is the discount factor. Similarly, the discounted and average cost problems are to select a sequence of policies γ_t, $t = 0, 1, \ldots$ satisfying (1) which respectively minimize $E\{\Sigma_{t=0}^{\infty} \beta^t r[s(t), u(t)] \mid p^0\}$ and $\limsup_{n \to \infty} \frac{1}{n+1} E\{\Sigma_{t=0}^{n} r[s(t), u(t)] \mid p^0\}$, where for the discounted cost case $\beta < 1$.

It has been shown that the partially observed Markov decision problem can be reformulated as the following completely observed Markov decision problem [1,7,8]. Let $\{x(t), t = 0, 1, \ldots\}$ and $\{\delta(t), t = 0, 1, \ldots\}$ be the (completely observed) state and control processes related by the stochastic difference equation $x(t+1) = \lambda\ [z(t+1), x(t), \delta(t)]$, $x(0) = p^0$, where $\lambda{:}S \times \Omega \times D \to \Omega$ has jth component $\lambda_j(k,x,y) = q_{jk}(y)\ \Sigma_i p_{ij}(y) x_i / \sigma(k,x,y) = P(s(t+1) = j \mid x(t) = x,\ z(t+1) = k,\ u(t) = y)$ and where $\sigma(k,x,y) = \Sigma_i\ \Sigma_j\ q_{jk}\ (y) p_{ij}\ (y) x_i = P(z(t+1) = k \mid\ x(t) = x,\ u(t) = y)$. The vector $x(t)$ can be thought of as the probability density of $s(t)$ on S conditioned on all past decisions and the past and present sample path of the observation process. Further interpretations of these terms can be found in [8,10]. It is sufficient [1,7] to consider policies of the form

$$\delta(t) = \gamma_t\,[x(t)]. \tag{2}$$

The criterion for the equivalent, finite horizon problem is $E\{\Sigma_{t=0}^{n-1} x(t) r(y) + x\ (n) r_0 \mid x(0)\}$, where $r(y) = \{r(i,y)\}$, $r_0 = \{r_0(i)\}$, and $xr = \Sigma_i x_i r_i$. The discounted and average cost criteria are modified appropriately; the equivalent problem is to select a sequence of policies satisfying (2) which minimizes the desired criterion.

The replacement problem is now described in the context of [4]. Define $h(x,y,\mathrm{v}) = xr(y) + \beta\ \Sigma_k\ \sigma(k,x,y)\, \mathrm{v}\,[\lambda(k,x,y)]$, where $\mathrm{v} \in \mathrm{V} = \{\mathrm{v} \in R^{\Omega}{:}\ 0 \leqslant \mathrm{v}(x) \text{ for all } x\}$ and where $\beta = 1$ for the average cost case. The scalar $h(x,y,\mathrm{v})$ can be interpreted as the expected cost to be accrued from time t until the terminal time, given that $x(t) = x$, $u(t) = y$, and if $x(t+1) = \lambda(k,x,y)$, then the expected cost to be accrued from time $t+1$ until the terminal time is $\mathrm{v}[\lambda(k,x,y)]$. Let Δ be the set of all functions $\delta{:}\Omega \to D$, i.e., $\Delta = D^{\Omega}$. It is desired to determine a subset of Δ, designated as Δ^* and called the set of *structured* policies, which will always contain optimal policies. The following definitions are preliminary to the definition of the Δ^* which will be examined throughout the remainder of this paper.

DEFINITION 1. (i) Let K be the set of all subsets of S such that if $i \in K \in \mathrm{K}$ and $i \leqslant i'$, then $i' \in K$.

(ii) Let the partial ordering $\prec$ on Ω be defined as $x \prec x'$ if and only if $xI_K \leqslant x'I_K$ for all $K \in \mathrm{K}$, where $xI_K = \Sigma_i x_i I_K(i)$ and where I_K is the indicator function of the set K, i.e., $I_K(i) = 0$ if $i \notin K$ and $I_K(i) = 1$ if $i \notin K$. Note that this partial ordering is equivalent to stochastic dominance.

For the case where the controller is not allowed completely informative on-line data of the system's state and has only the process $\{x(t),\ t = 0, 1, \ldots\}$ available on which to base a decision, it will be shown that if $x \prec x'$, then state x' is not more desirable than state x.

The restricted set of policies is now defined and will be shown in the following section to always contain an optimal policy under reasonable assumptions. This set generalizes the usual notion [2] of (and will also be referred to as) the set of all *control-limit* policies.

DEFINITION 2. $\Delta^* = \{\delta \in \Delta{:}x \prec x' \text{ implies } \delta(x) \leqslant \delta(x')\}$.

It will also be shown that control-limit policies will induce structural properties of their associated expected cost functions. The set of all such cost functions is now defined.

DEFINITION 3. Define $V^* = \{v \epsilon V : v \text{ is concave, and } x \prec x' \text{ implies } v(x) \leqslant v(x')\}$.

We complete this section with the following definitions. The optimal expected cost to be accrued between time t and the terminal time n is $f_{n-t} = \inf_{\pi} (H_{\delta_t} \times \ldots \times H_{\delta_{n-1}} u_0)$, where $u_0(x) = xr_0$, $\pi = (\delta_0, \ldots, \delta_{n-1})$, and $[H_\delta v](x) = h[x, \delta(x), v]$. The optimal discounted cost accrued over the infinite horizon is $f = \inf_{\pi} \lim_{n \to \infty} (H_{\delta_t} \times \ldots \times H_{\delta_{n-1}} u_0)$.

3. STRUCTURE OF OPTIMAL STRATEGIES AND COST FUNCTIONS FOR THE COMPLETELY OBSERVED AND COMPLETELY UNOBSERVED CASES

In this section two special cases of the general replacement problem described above are considered, the completely observed case (where $q_{jk}(y)$ is the Kronecker delta for all y) and the completely unobserved case (where $q_{jk}(y)$ is independent of j for all k and y, i.e., $q_{jk}(y) = q_k(y)$). The completely observed case is a direct generalization of results in [3] and has been examined under slightly different assumptions in [9]. The completely unobserved case is a generalization of a slightly different version of Example 2 [5, p. 130]. It will be shown that for both cases an optimal strategy can be found which is composed of policies in Δ^*, thus inducing optimal expected cost functions which are members of V^*.

Two assumptions are now stated, following a preliminary definition, which will be shown to be sufficient for the existence of optimal strategies composed of policies in Δ^* for both special cases.

DEFINITION 4. Let $F = \{\xi \epsilon R^S : i \leqslant i' \text{ implies } \xi(i) \leqslant \xi(i')\}$, the set of all real-valued functions on S which are increasing with respect to the partial ordering $\prec$.

ASSUMPTION 1. The cost functions $r(\cdot, y)$, for all $y \in D$, $r_0(\cdot)$, and $r(\cdot, 0) - r(\cdot, 1)$ are members of F.

ASSUMPTION 2. $PI_K \in F$ for all $K \in K$, where the ith element ($i \in S$) of PI_K is $\Sigma_{j \in S} p_{ij} I_K(j)$.

Assumption 1 states that operating and replacement costs and their difference increase as a function of system state deterioration; similarly, terminal cost (for the finite horizon problem) increases as a function of system state deterioration. Assumption 2 is a generalization of the increasing failure rate assumption. Both of these assumptions are direct generalizations of those made by Derman in [3] for the scalar, completely observed case.

THE COMPLETELY UNOBSERVED CASE

The case where the controller receives no on-line data of any value with regard to the actual state of the system is now examined. For this case $h(x, 0, v) = xr(0) + \beta v(xP)$ and $h(x, 1, v) = xr(1) + \beta v(e_0)$, where the jth element of xP is $\Sigma_i p_{ij} x_i$ and where $e_0 \epsilon \Omega$ has 1 as its 0th element. Two preliminary results are now presented.

LEMMA 1. For $x,x'\epsilon\Omega$, $x \prec x'$ if and only if $x\,\xi \leqslant x'\xi$ for all $\xi\epsilon$F.

PROOF: Clearly, $I_K \in$F for all $K \in$K. Thus, $x\xi \leqslant x'\xi$ for all $\xi \in$F trivially implies $x \prec x'$.

Conversely, it is easily shown that for each $\xi\epsilon$F there exists a non-negative sequence $\{\alpha_K\}$ such that $\xi = \sum_{K\epsilon K} \alpha_K I_K$. It then follows that $x \prec x'$ implies the equality-inequality chain $x\xi = \Sigma_K \alpha_K x I_K \leqslant \Sigma_K \alpha_K x' I_K = x'\xi$.

Q.E.D.

Lemma 1 and Assumption 2 imply the following corollary. (See also Theorem 2.1b in [2].)

COROLLARY 1. If $x \prec x'$, then $\lambda(x,y) \prec \lambda(x',y)$, for each $y\epsilon D$, where $\lambda(k,x,y) = \lambda(x,y)$, is independent of k for the completely unobserved case.

Two propositions are now presented which are concerned with the structural properties of the function $h(\cdot,y,\mathrm{v})$ for the completely unobserved case.

PROPOSITION 1. If $\mathrm{v}\epsilon V^*$, then $h(\cdot,y,\mathrm{v})\epsilon V^*$ for each $y\epsilon D$.

PROOF: This result follows directly from the definitions, Assumptions 1 and 2, Lemma 1 , and Corollary 1.

PROPOSITION 2. Assume $\mathrm{v}\epsilon V^*$ and $x \prec x'$. (a) If $h(x,0,\mathrm{v}) - h(x,1,\mathrm{v}) \geqslant 0$, then $h(x',0,\mathrm{v}) - h(x',1,\mathrm{v}) \geqslant 0$. (b) If $h(x',1,\mathrm{v}) - h(x',0,\mathrm{v}) \geqslant 0$, then $h(x,1,\mathrm{v}) - h(x,0,\mathrm{v}) \geqslant 0$.

PROOF: This result also follows directly, where we note

$$h(x,0,\mathrm{v}) - h(x,1,\mathrm{v}) = x[r(0) - r(1)] + \beta\{\mathrm{v}[\lambda(x,0)] - \mathrm{v}(e_0)\}. \qquad \text{Q.E.D.}$$

Two additional propositions are now presented which will be useful in the proofs of the main results. Define the operator A as $A\mathrm{v} = \inf_{\delta} H_\delta \mathrm{v}$.

PROPOSITION 3. $A: V^* \rightarrow V^*$.

PROOF: The finiteness of D implies the existence of a $\delta\in\Delta$ such that $w(x) = \min_{y\in D} h(x,y,\mathrm{v}) = h[x, \delta(x), \mathrm{v}]$. Since the minimum of concave functions is concave, w is concave by Proposition 1. Letting $x \prec x'$, the result follows by the equality-inequality chain $w(x) \leqslant h(x, \delta(x'), \mathrm{v}) \leqslant h(x', \delta(x'), v) = w(x')$.

Q.E.D.

PROPOSITION 4. If $v \in V^*$, then there exists a $\delta\in\Delta^*$ such that $H_\delta \mathrm{v} = A\mathrm{v}$.

PROOF: Assume that $\delta\in\Delta$ satisfies $H_\delta\mathrm{v} = A\mathrm{v}$ but that $\delta\notin\Delta^*$. Let $x \prec x'$, $\delta(x') = 0$ and $\delta(x) = 1$. By Proposition 2, since $h(x,1,\mathrm{v}) \leqslant h(x,0,\mathrm{v})$ it follows that $h(x',1,\mathrm{v}) \leqslant h(x',0,\mathrm{v})$ and hence since $h(x',1,\mathrm{v}) \geqslant h(x',0,\mathrm{v})$, $h(x',0,\mathrm{v}) = h(x',1,\mathrm{v})$. Thus, δ could have been chosen so that $\delta(x') = 1$, and the result holds.

Q.E.D.

The main results for the finite horizon and discounted cost problem can now be presented.

THEOREM 1. Consider the finite horizon problem. It follows that for all n, the optimal expected cost function $f_n \in V^*$ and that there exists a sequence of control-limit policies which is simultaneously for all t optimal for stages t to n.

PROOF: This result is an application of Theorem 1 in [4] to the completely unobserved case, which requires Propositions 3 and 4 for satisfaction of two of the six hypotheses specified in [4]. The remaining hypotheses follow from the definitions.

Q.E.D.

Theorem 1 guarantees that the optimal expected cost to be accrued over a finite horizon of length n is concave and nondecreasing with respect to $<$ in the a priori probability. Also, in searching for an optimal strategy, it is sufficient to only examine strategies composed of control-limit policies.

THEOREM 2. Consider the discounted cost problem. It follows that the optimal expected discounted cost function $f \in V^*$ and that there exists an optimal stationary control-limit strategy.

PROOF: The above result is a specialization of Theorem 2 and Corollary 1 and [4] to the replacement problem. These results require that nine hypotheses, specified in [4], be satisfied, the first six of which were required for Theorem 1 above. If $\Omega^* = \Omega$, $\bar{r} = \max_{i,y} r(i,y)$; $V^0 = \{v \in V^* : v(x) \leqslant \bar{r}/(1-\beta)$, for all $x \in \Omega^*\}$, $N = 1$ and $M = 2\bar{r}/(1-\beta)$, then the remaining three hypotheses are easily verifiable (where the notation in this proof corresponds to the notation in [4].)

The above result states that the optimal expected discounted cost is also concave and nondecreasing with respect to $<$ in the a priori probability and that there exists an optimal stationary strategy generated by a control-limit policy.

The average cost case is now examined. Let f_β designate the optimal expected cost function for the discounted cost case with discount factor β. We appeal to results in [5, pp. 141-150] which require Ω^* to be countable. Redefine $\Omega^* = \{p_{i.}^{(k)}$: $i \in S$, $k = 0, 1, \ldots\}$, where $p_i^{(k)} = \{p_{ij}^{(k)}\}$ and $p_{ij}^{(k)} = P\{s(t+k) = j \mid s(t) = i\}$. Let $e_i \in \Omega$ have 1 as its ith element. Note that the results of Theorem 2 hold for this Ω^* since $e_i \in \Omega^*$ for all $i \in S$ and $\lambda(x,y) \in \Omega^*$ for all $x \in \Omega^*$ and $y \in D$.

THEOREM 3. Consider the average cost case. Then, there exists a function $\bar{f} \in V^*$ which is the optimal expected average cost function and a constant $\bar{g}$ such that $\bar{g} + \bar{f}(x) = \min \{xr(0) + \bar{f}[\lambda(x,0)], xr(1) + \bar{f}(e_0)\}$ and there exists an optimal stationary control-limit strategy which causes the above minimum to be attained.

PROOF: Note that $|f_\beta(x) - f_\beta(e_0)| \leqslant \bar{r}$ for all β and x since $f\beta(x) - f_\beta(e_0) \geqslant 0$ for all x (clearly, $e_0 \prec x$ for all x) and $f_\beta(x) \leqslant \bar{r} + \beta f_\beta(e_0) \leqslant \bar{r} + f_\beta(e_0)$. Theorem 6.18 in [5, p. 146] is then satisfied. It follows from Theorem 6.18 (ii) in [5] that $\bar{f} \in V^*$ since $f_\beta \in V^*$ implies $f_\beta(\cdot) - f_\beta(e_0) \in V^*$ and since V^* is complete. The result of Proposition 4 implies that the minimizing policy can be chosen from Δ^*.

Q.E.D.

Thus, it is again sufficient to examine control-limit policies in search of an optimal strategy and the optimal expected average cost function is concave with (with respect to Ω) and nondecreasing in $\prec$.

The completely unobserved problem is now reformulated into a form compatible with the computational algorithm presented in [9]. Let state t for the countable state reformulated problem be equivalent to state $e_0 P^t$ and assume that the a priori probability $p^0 = e_0 P^t$ for some t. The transition probabilities for the reformulated problem are then $p'_{tt+1}(0) = 1$ and $p'_{t0}(1) = 1$. Define $r'(t,y) = e_0 P^t r(y)$ and $r'_0(t) = e_0 P^t r_0$. Note that Assumptions 1 and 2 hold for the reformulated problem, where elements $K' \in \mathrm{K}'$ (analogous to K and K in the original problem) are of the form $K' = \{t, t+1, \ldots\}$. Since $\{p_{ts}\}$ is such that $p_{ts} = 0$ if $s \leqslant t$, then all of the hypotheses for the use of the algorithm presented in [9] are satisfied and hence the algorithm is applicable for the completely unobserved case.

THE COMPLETELY OBSERVED CASE.

The case where the controller knows the present state of the system perfectly at each decision epoch is now briefly examined and follows the outline (and leads to the identical results) for the completely unobserved case.

For the completely observed case, $h(x,0,\mathrm{v}) = xr\ (0) + \beta x P \mathrm{v}'$ and $h(x,1,\mathrm{v}) = xr(1) + \beta \mathrm{v}_0$, where $\mathrm{v}_i = \mathrm{v}(e_i)$. Note that $\mathrm{v} \in V^*$ implies $\mathrm{v}' \in \mathrm{F}$ and that $h'(.,y,\mathrm{v}) \in \mathrm{F}$ implies $h(\cdot,y,\mathrm{v}) \in \mathrm{V}^*$ due to the linearity of $h(\cdot,y,\mathrm{v})$, where $h'(i,y,\mathrm{v}) = h(e_i,y,\mathrm{v})$ (and where the notion of concavity for functions in V^* is no longer meaningful). Note also that $PI_K \in \mathrm{F}$ for all $K \in \mathrm{K}$ is equivalent to $P\xi \in \mathrm{F}$ for all $\xi \in \mathrm{F}$ (which is a generalization of the equivalence of Conditions A and B in [3]). These results, and the fact that $f_n(x) = \Sigma_i x_i f_n(e_i)$ for all n for the completely observed case [1], can easily be used to imply that Propositions 1 through 4 and Theorems 1 through 3 also hold for the completely observed case, subject to obvious modifications to Ω^* and (3). Such results are similar in form (although not in approach) to results presented in [9]. Furthermore, if P is such that $p_{ij} = 0$ unless $i \leqslant j$, then the computational algorithm presented in [9] is applicable to the completely observed case.

4. CONCLUSIONS

This paper has examined the structural properties of an optimal strategy and of the optimal expected cost function for two special cases of a replacement problem subject to Markov deterioration. Under mild assumptions, it has been shown for these two cases that there exits optimal strategies with a generalized control-limit structure. Under additional assumptions these structured strategy results have implied that a recently developed, simplified computational approach is valid for both cases. Interest in these two special cases has resulted from the fact that their optimal cost functions represent bounds on the optimal cost function for the general case, which unfortunately does not necessarily possess an optimal structured strategy.

REFERENCES

[1] Astrom, K.J., "Optimal Control of Markov Processes with Incomplete State Information," Journal of Mathematical Analysis and Applications, Vol. 10, pp. 174-205 (1965).

[2] Barlow, R.E. and F. Proschan, "Theory of Maintained Systems: Distribution of Time to First System Failure," Mathematics of Operations Research, Vol. 1, pp. 32-42 (1976).

[3] Derman, C., "On Optimal Replacement Rules When Changes of State are Markovian," in *Mathematical Optimization Techniques*, R. Bellman (ed.), (U. of Calif. Press, Berkeley, 1963).

[4] Porteus, E.L., "On the Optimality of Structured Policies in Countable Stage Decision Processes," Management Science, Vol. 22, pp. 148-157 (1975).

[5] Ross, S.M., *Applied Probability Models with Optimization Applications*, (Holden-Day, San Francisco, 1970).

[6] Sandell, N.R., Jr., "Control of Finite-State, Finite-Memory Stochastic Systems," Sc.D. thesis, E.E. Dept. MIT, Cambridge, Mass. (1974).

[7] Sawaragi, K. and T. Yoshikawa, "Discrete-Time Markovian Decision Processes with Incomplete State Information," Annals of Mathematical Statistics, Vol. 41, pp. 78-86 (1970).

[8] Smallwood, R.D. and E.J. Sondik, "The Optimal Control of Partially Observable Markov Processes Over a Finite Horizon," Operations Research, Vol. 21, pp. 1300-1322 (1973).

[9] Wang, R.C., "Computing Optimal Replacement Policies-Discrete States," Mountain Bell Tech. Report, Denver, Colorado (May 1975).

[10] White, C.C., "Procedures for the Solution of a Finite-Horizon, Partially Observed, Semi-Markov Optimization Problem," Operations Research, Vol. 24, pp. 348-358 (1976).

[11] White, C.C., "Optimal Control-Limit Strategies for a Partially Observed Replacement Problem, "International Journal of System Science, Vol. 10, pp. 321-331 (1979).

SURVIVAL DISTRIBUTIONS IN CROSSING FIELDS CONTAINING CLUSTERS OF MINES WITH POSSIBLE DETECTION AND UNCERTAIN ACTIVATION OR KILL*

S. Zacks

Case Western Reserve University
Cleveland, Ohio

ABSTRACT

The present paper presents an algorithm for the exact determination of survival distributions in crossing mine fields. The model under consideration considers clusters of mines, scattered at random in the field around specified aim points. The scatter distributions of the various clusters are assumed to be known. The encounter process allows for a possible detection and destruction of the mines, for inactivation of the mines and for the possibility that an activated mine will not destroy the object. Recursive formulae for the determination of the survival probabilities of each object (tank) in a column of n crossing at the same path are given. The distribution of the number of survivors out of n objects in a column is also determined. Numerical examples are given.

1. INTRODUCTION

In the present paper we develop methods for the exact numerical determination of the survival probabilities of objects (targets) crossing a field containing randomly scattered mines. The model under consideration refers to cases in which the absorption points are randomly scattered over the field in one or several clusters. Each cluster is characterized by some bivariate distribution of the mines around a center (aimpoint). More specifically we consider clusters which are distributed either as bivariate normal or uniformly over rectangular domains. The objects cross the field in columns along predetermined breaching paths. The movement of the objects through the field is in a continuous manner (like that of vehicular or tank targets). The mines may be detected by the objects and destroyed. On the other hand, if a mine is not detected it may or may not be activated. If it is not activated in a specific encounter it may be activated in following encounters. Defective mines (duds) which can never be activated play no significant role in the determination of the survival probabilities. We have to know the proportion of defective mines (duds) only in order to determine correctly the distributions of the anticipated number of active mines in the various possible crossing paths.

The specific details of the probabilistic model, as related to the structure of the field, is described in Section 2.

There are several papers in the available literature [1,2,3,4] which study similar models. In most of these papers the results are based on computer simulation. Both the location of

*Supported by the Office of Naval Research under Contract NR 00014-75-C-0529, Project NR 042-276, at Case Western Reserve University.

mines and the results of the encounters are determined by Monte Carlo methods. Although such an approach can provide estimates of the required survival probabilities it suffers from the following deficiencies: (i) the amount of needed computer time is excessive; (ii) the estimates are generally not accurate enough; (iii) lack of theoretical basis. In 1966 and 1973 Zacks and Goldfarb [6] and Parsons [5] provided analytical methods for the derivation of formulae for the *exact* determination of the survival probabilities. They treated, however, a model which is too restricted. In the present paper we provided the formulae and the algorithms for the determination of survival distribution *exactly*. The methods illustrated in the example of Section 6 are very efficient, and require on a fast computer (like the UNIVAC 1108) only a few seconds for the exact numerical solution. Fortran programs, according to which the numerical results have been computed, are available and can be obtained upon request. In the following theoretical development a mine is called an "absorption-point" and the object is called a "particle." This allows application of the theory to more general areas.

2. STRUCTURAL AND PROBABILITY MODEL OF FIELDS

In the present section we provide specific models (sets of assumptions) concerning the structure of the absorption field; the scatter distributions of the absorption points; and the crossing procedures.

A. The field structure

We assume that the absorption field is rectangular of length L and width W. We fix the origin of the coordinates system at the center of the field. Crossing paths are along straight lines parallel to the y-axis, of width DL. The center of the crossing path is at a point $(B,0)$. The point $(B, -W/2)$ is called the *breach-point*.

The absorption points are scattered over the field at random, in clusters of N points distributed according to specified bivariate distributions centered at given *aimpoints*. Let k be the number of clusters in the field. The coordinates of the *aimpoints* are $(\xi^{(i)}, \eta^{(i)})$, $i = 1, \dots, k$. In some applications the aimpoints are located along straights lines called *aimpoint rows*. Let r be the number of aimpoint rows. There are m_i aimpoints in the i-th row and $\sum_{i=1}^{r} m_k = k$. In this case the coordinates of the aimpoints can be determined according to the rows to which they belong. Typical clusters of absorption points in a field can be described as in Figure 2, when the points have bivariate normal distributions along aimpoint rows which are not parallel to the x-axis.

B. The encounter model

In the present study we consider a general model based on the following assumptions.

(i) Absorption points act independently of each other.

(ii) In each encounter, independently of the past history, an absorption point can be detected and destroyed. The probability of detection p_{det}, remains fixed throughout all the encounters.

(iii) An undetected absorption point may be activated with probability, p_{act}, independently of any previous event. If a point is activated it is destroyed.

(iv) A particle activating an absorption point is either absorbed or survives. The probability of absorption at each activation is p_k. In each case absorption or survival are independent of the previous events.

In the previous studies of Zacks and Goldfarb [6] and Parsons [5] no detection was allowed and if a point was activated the particle was always destroyed.

3. SCATTER DISTRIBUTION MODELS

In each cluster the N points are distributed identically and independently according to some specified distribution. The scatter distributions do not have to be the same at different clusters. In the present section we discuss the bivariate normal and the uniform scatter distributions. The methodology remains the same if other distributions are chosen.

Let (X_{ij}, Y_{ij}); $j = 1, \dots, N_i$ and $i = 1, \dots, k$ designate the coordinates of the j-th absorption point at the i-th cluster. The (x,y) coordinates system is the one with origin at the center of the field and axes parallel to the rectangular sides of the field (Figure 1). The absorption point may be scattered over the field according to another coordinate system (x', y'), which can be obtained by a rotation of the (x,y) system. We introduce the (x',y') system since the aimpoints may be scattered (for example, from the air) along aimpoint rows which are not parallel to the (x,y) axes. We denote by θ the angle between the x and x' axes. It is well known that the (x,y) coordinates can be obtained from the (x',y') coordinates by the orthogonal transformation

(3.1)
$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix}.$$

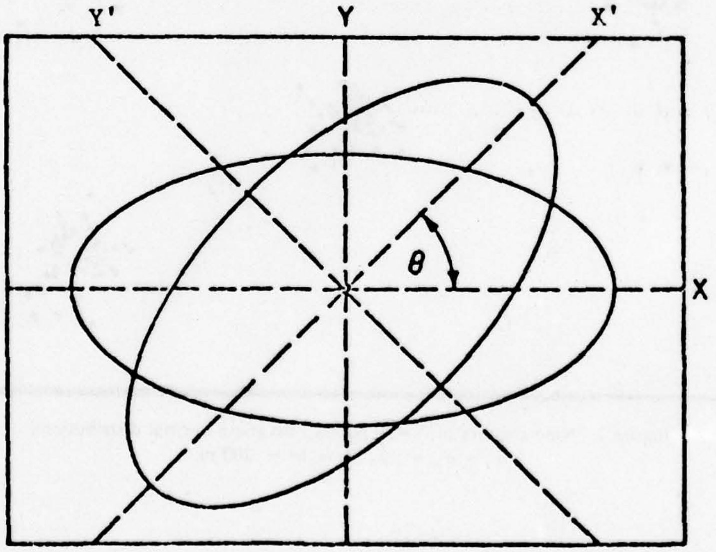


Figure 1. Ellipsoids of concentration of scatter distributions.

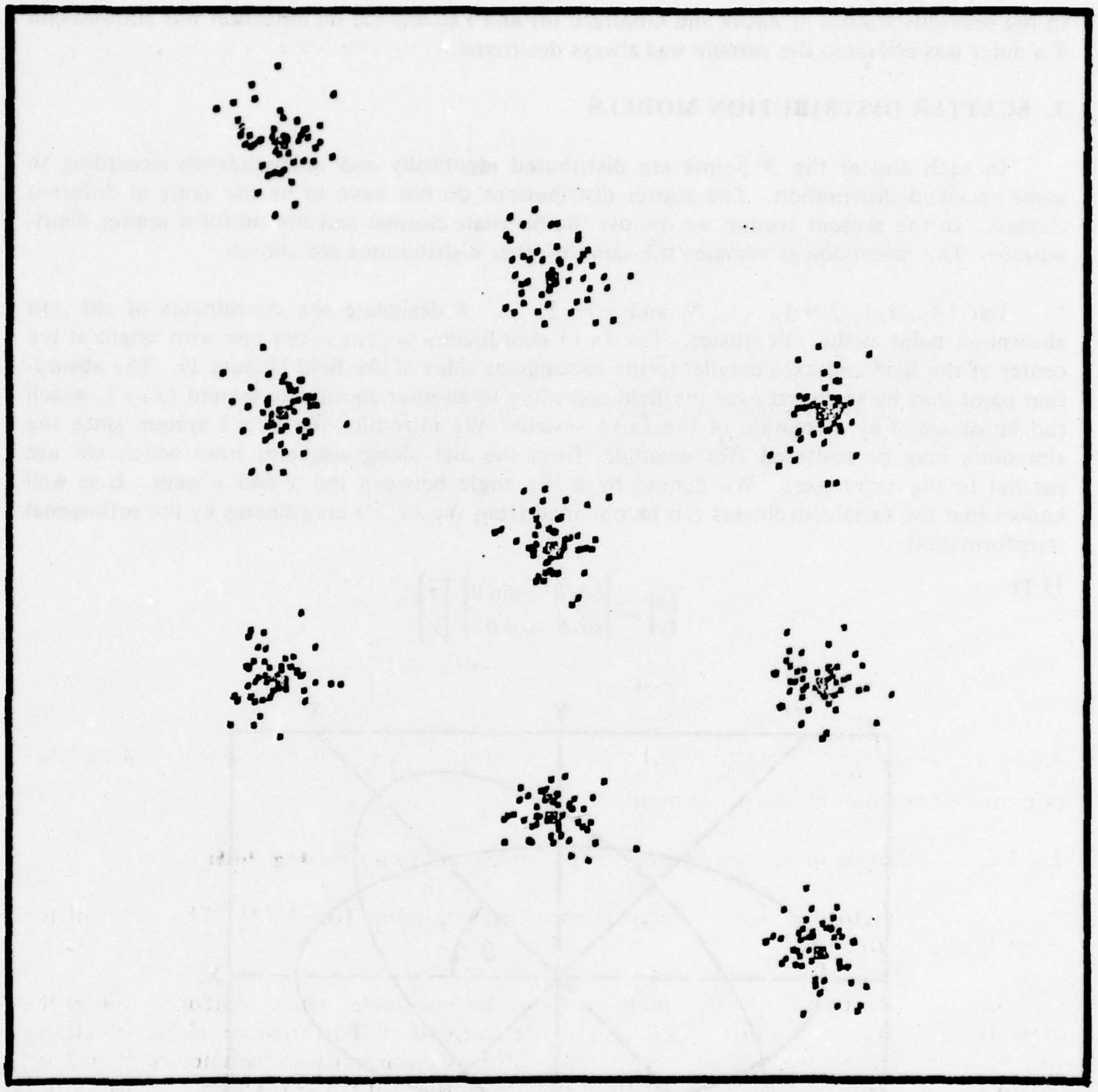


Figure 2. Nine clusters of $N = 50$ points. Bivariate normal distributions $\sigma_x = \sigma_y = 5\text{m}$. $L = W = 200$ m.

3.1 The Bivariate Normal Model

It is assumed that the scatter of points in a given cluster, according to the (x',y') system, follows a bivariate normal distribution whose mean is the aimpoint (ξ',η'). Furthermore, the random variables x' and y' are independent with variances σ_x^2 and σ_y^2, respectively. In terms of the (x,y) coordinate system, the points (x,y) are normally distributed with mean vector (ξ,η) and a covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \tag{3.2}$$

where

$$\begin{aligned} \sigma_{11} &= \sigma_x^2 \cos^2\theta + \sigma_y^2 \sin^2\theta \\ \sigma_{12} &= \frac{1}{2}(\sigma_x^2 - \sigma_y^2)\sin(2\theta) \\ \sigma_{22} &= \sigma_x^2 \sin^2\theta + \sigma_y^2 \cos^2\theta \end{aligned} \tag{3.3}$$

If ρ denotes the coefficient of correlation between x and y, i.e., $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$, then the joint probability density of (x,y) is

$$\begin{aligned} &f(x,y \mid \xi,\eta,\sigma_{11},\sigma_{12},\sigma_{22}) \\ &= \frac{1}{2\pi}\,\frac{1}{\sqrt{\sigma_{11}\,\sigma_{22}(1-\rho^2)}}\exp\left\{-\frac{1}{2\sigma_{11}}(x-\xi)^2 \right. \\ &\qquad \left. -\frac{1}{2\sigma_{22}(1-\rho^2)}[y-\eta-\beta(x-\xi)]^2\right\}, \end{aligned} \tag{3.4}$$

where $\beta = \rho\sqrt{\frac{\sigma_{22}}{\sigma_{11}}}$. As mentioned earlier, the parameters of this bivariate normal distribution may differ from one cluster to another.

3.2 The Distribution of the Number of Absorption Points in a Crossing Path

Consider a crossing path centered at the breaching point $(B,-W/2)$. The width of the crossing path is DL.

Let J_i be the number of absorption points of the i-th cluster which (randomly) fall in the crossing path. We assume that $J_1, J_2, \ldots, J_k$ are *independent.* Furthermore, if the i-th cluster has N_i absorption points then the distributions of J_i is *binominal* with parameters N_i and ψ_i; where ψ_i is the probability that a point from the i-th cluster will fall in the path. We call ψ_i the *coverage probability* of the i-th cluster. formulae for ψ_i in the bivariate normal case are obtained as follows.

Let $\phi(u)$ and $\Phi(u)$ denote the p.d.f. and the cumulative distribution function (c.d.f.) of the standard (univariate) normal distribution. Then

$$\psi_i = \frac{1}{\sqrt{\sigma_{11}}} \int_{B-\frac{DL}{2}}^{B+\frac{DL}{2}} \phi\left(\frac{x-\xi_i}{\sqrt{\sigma_{11}}}\right) \left\{\frac{1}{\sqrt{\sigma_{22}(1-\rho^2)}} \int_{\frac{-W}{2}}^{\frac{W}{2}} \phi\left(\frac{y-\eta_i-\beta(x-\xi_i)}{\sqrt{\sigma_{22}(1-\rho^2)}}\right) dy\right\} dx$$

$$(3.5) \qquad = \frac{1}{\sqrt{\sigma_{11}}} \int_{B-\frac{DL}{2}}^{B+\frac{DL}{2}} \phi\left(\frac{x-\xi_i}{\sqrt{\sigma_{11}}}\right) \left[\Phi\left(\frac{\frac{W}{2} - \eta_i - \beta(x-\xi_i)}{\sqrt{\sigma_{22}(1-\rho^2)}}\right)\right.$$

$$\left. + \Phi\left(\frac{\frac{W}{2} + \eta_i + \beta(x-\xi_i)}{\sqrt{\sigma_{22}(1-\rho^2)}}\right)\right] dx - \left[\Phi\left(\frac{B + \frac{DL}{2} - \xi_i}{\sqrt{\sigma_{11}}}\right) - \Phi\left(\frac{B - \frac{DL}{2} - \xi_i}{\sqrt{\sigma_{11}}}\right)\right].$$

Since DL is generally substantially smaller than $\sqrt{\sigma_{11}}$ a good approximation for (3.5) is obtained by

$$(3.6) \qquad \psi_i \cong \left\{\Phi\left[\frac{\frac{W}{2} - \eta_i - \beta(B - \xi_i)}{\sqrt{\sigma_{22}(1-\rho^2)}}\right]\right.$$

$$\left. + \Phi\left(\frac{\frac{W}{2} + \eta_i + \beta(B - \xi_i)}{\sqrt{\sigma_{22}(1-\rho^2)}}\right) - 1\right\} \cdot \left\{\Phi\left(\frac{B + \frac{DL}{2} - \xi_i}{\sqrt{\sigma_{11}}}\right) - \Phi\left(\frac{B - \frac{DL}{2} - \xi_i}{\sqrt{\sigma_{11}}}\right)\right\}.$$

Alternatively, one can apply any program which computes bivariate normal probabilities of rectangles for the determination of ψ_i.

3.3 The Distribution of J

As mentioned earlier, the number of absorption points, J, in a crossing path is a sum of k independent binomial random variables $J_1, \ldots, J_k$. These binomial random variables have different coverage probability parameters, $\psi_i (i = 1, \ldots, k)$. Hence, the distribution of J is not binomial. One can determine the formula of the probability distribution function of J by determining first its generating function. We obtain that, if $\omega_i = \psi_i/(1-\psi_i)$, $i = 1, \ldots, k$, and $Q = \prod_{i=1}^{k}(1-\psi_i)^{N_i}$ then

$$(3.7) \qquad \begin{aligned} &P\{J_1 + \cdots + J_k = r \mid \psi_1, \ldots, \psi_k\} \\ &= Q \sum_{\{j_1+\cdots+j_k=r\}} \cdots \sum \exp\left\{\sum_{i=1}^{k}\left[\log\binom{N_i}{j_i} + j_i \log \omega_i\right]\right\}, \end{aligned}$$

$r = 0, 1, \ldots, N_1 + \cdots + N_k$.

When N_i is large and ψ_i small one would encounter difficulties in computing these probabilities exactly. In these cases the Poisson approximation is generally applicable. The distribution of $J_1 + \cdots + J_k$ is well approximated, when ψ_i are small and N_i are large, by the Poisson distribution with mean $\lambda = \sum_{i=1}^{k} N_i\psi_i$.

4. THE SURVIVAL PROBABILITIES OF PARTICLES

In the present section we develop the formulae of the survival probabilities of each one of n particles crossing consecutively in a given path. As before, let J denote the number of absorption points in the breaching path. We assume here that all these points are non-defective. Let s denote the probability of survival when the object passes in the vicinity of an absorption point. From the assumption of the model

$$s = p_{\text{det}} + (1 - p_{\text{det}})\,[(1 - p_{act}) + p_{act}(1 - p_k)]. \tag{4.1}$$

Let S_n denote the event that the n-th object to cross a path survives. Since the encounters are independent it is obvious that

$$P[S_1 | J] = s^J \tag{4.2}$$

The survival probability of the first particle is then the expected value of the conditional probability, i.e.,

$$P[S_1] = E\{s^J\}. \tag{4.3}$$

We develop now general recursive formulae for the determination of survival probabilities $P[S_n]$ for all $n \geqq 2$.

Let M_n, $n = 1, 2, \ldots$ be a random variable designating the total number of absorption points destroyed (detected or activated) during the first n attempts to cross in a given path. Obviously,

$$P[S_n | M_{n-1}, J] = s^{J-M_{n-1}}, \quad \text{all} \quad n = 1, 2, \ldots \tag{4.4}$$

where $M_0 \equiv 0$. Thus, if we determine the conditional distribution of M_i given J, for all $i = 1, 2, \ldots$ we can obtain from (4.4) the conditional probability of S_n given J, namely

$$\begin{aligned} P[S_n | J] &= \sum_{r=0}^{J} P[S_n | M_{n-1} = r, J] \cdot P[M_{n-1} = r | J] \\ &= \sum_{r=0}^{J} s^{J-r}\, P[M_{n-1} = r | J]. \end{aligned} \tag{4.5}$$

For each $n = 1, 2, \ldots$ let Δ_n denote the number of absorption points destroyed in the n-th attempt to cross the path. Since $M_n = M_{n-1} + \Delta_n$ for $n = 1, 2, \ldots$, we obtain the (recursive) relationship, for all $r = 0, =1, \ldots, J$ and every $n = 1, 2, \ldots$

$$P[M_n = r | J] = \sum_{k=0}^{r} P[\Delta_n = k | M_{n-1} = r-k, J] \cdot P[M_{n-1} = r-k | J]. \tag{4.6}$$

Moreover, for every $n = 1, 2, \ldots ;\ k = 0, \ldots, r$;

$$P[\Delta_n = k | M_{n-1} = r-k, J] = P[M_1 = k | J - r+k]. \tag{4.7}$$

Hence, from (4.6) and (4.7) we obtain

$$P[M_n = r | J] = \sum_{k=0}^{r} P[M_1 = k | J-r+k]\; P[M_{n-1} = r-k | J]. \tag{4.8}$$

Thus, if we develop explicit formula of $P[M_1 = r | J = j]$ for all $j = 0, 1, \ldots$ and $r = 0, 1, \ldots, j$ then we can obtain, by repeatedly applying (4.8), the values of $P[M_n = r | J = j]$ for all $n \geqslant 2$, $r = 0, \ldots, j$ and $j = 0, 1, \ldots$. This explicit formula will be derived in the sequel. We return now back to (4.5). The conditional probability of S_2 given J is

$$P[S_2|J] = \sum_{r=0}^{J} S^{J-r}\, P[M_1 = r|J]. \tag{4.9}$$

For $n \geqslant 3$ we obtain from (4.5) and (4.8)

$$\begin{aligned} P[S_n|J] &= \sum_{r=0}^{J} s^{J-r} \sum_{k=0}^{r} P[M_{n-2} = k|J]\, P[M_1 = r-k|J-k] \\ &= s^J \sum_{k=0}^{J} P[M_{n-2} = k|J] \sum_{r=k}^{J} S^{-r}\, P[M_1 = r-k|J-k] \\ &= s^J \sum_{k=0}^{J} s^{-k}\, P[M_{n-2} = k|J] \sum_{i=0}^{J-k} s^{-i}\, P[M_1 = i|J-k]. \end{aligned} \tag{4.10}$$

We have thus established the recursive relation, for all $n \geqslant 3$,

$$P[S_n|J] = \sum_{r=0}^{J} P[M_{n-2} = r|J]\; P[S_2|J-r]. \tag{4.11}$$

As seen in formulae (4.8) and (4.9) the key to the recursive solution is in the explicit expression of $P[M_1 = r|J]$. We derive now the formula for these conditional probabilities. We notice first that no absorption point is destroyed in the first crossing if, and only if, there were no detections and no activations. Thus,

$$P[M_1 = 0|J] = w^J \tag{4.12}$$

where $w = (1-p_{\text{det}})\,(1-p_{act})$. Similarly, the event$\{M_1 = J\}$ is possible if, and only if, all the first J absorption points have been either detected or activated and the particles survived at least the first J-1 encounters. Thus,

$$P[M_1 = J|J] = (1-w)\,(s-w)^{J-1}. \tag{4.13}$$

In the other case (r points destroyed; $r = 1, \ldots, J-1$) we distinguish between two possibilities. The first one is that the particle has survived the r encounters, in which the points have been destroyed, and the other one is that it has been also destroyed. The probability of r absorption points destroyed and the particle surviving is $\binom{J}{r}(s-w)^r\, w^{J-r}$. On the other hand if the particle is also destroyed it could have been destroyed only after the first $r-1$ encounters. Accordingly, the probability of r absorption points destroyed and the particle destroyed too, is

$$(1-s)\,(s-w)^{r-1} \sum_{j=r-1}^{J-1} \binom{j}{r-1} w^{j-(r-1)} = \frac{1-s}{1-w}\left(\frac{s-w}{1-w}\right)^{r-1} G(J-r|w,r), \tag{4.14}$$

where $G(j|\psi,\nu)$ is the c.d.f. of the negative-binomial probability distribution, having a probability function

$$g(j|\psi,\nu) = \frac{\Gamma(j+\nu)}{\Gamma(j+1)\,\Gamma(\nu)}\,(1-\psi)^\nu\, \psi^j, \tag{4.15}$$

$j = 0, 1, \ldots; 0 < \psi < 1$ and $0 < \nu < \infty$. Accordingly, from (4.12)–(4.14) we obtain the general formula

$$P[M_1 = r|J] = \begin{cases} w^J & \text{, if } r = 0 \\ \binom{J}{r} w^{J-r}(s-w)^r + \\ \dfrac{1-s}{1-w}\left(\dfrac{s-w}{1-w}\right)^{r-1} G(J-r|w,r) & \text{, if } r = 1, \ldots, J-1 \\ (1-w\;(s-w)^{J-1} & \text{, if } r = J \end{cases} \tag{4.16}$$

It can be shown that $\sum_{r=0}^{J} P[M_1 = r|J] = 1$ for all $J = 0, 1, \ldots$. Thus, the computing algorithm should start with the computation of the values of $P[M_1 = r|J]$ for all $J = 0, 1, \ldots$ and $r = 0, 1, \ldots, J$. In the second stage of the computations the conditional probabilities $P[s|J]$ can be determined for each $J = 0, 1, \ldots$. These probabilities are then multiplied by the probabilities of J and summed over all $j = 0, 1, 2, \ldots$. This sum will provide the probability $p[S_2]$. In more progressive stages we compute $P[M_n = r|J]$ and $P[S_n|J]$ according to the previous recursive formulae and then determine the weighted average, the conditional probabilities $P[S_n|J]$.

We remark that since $M_n \geqq M_{n-1}$, with probability one, for all $n \geqq 1$, and since $0 < s < 1$,

$$P[S_{n+1}|J,M_n] = s^{J-M_n} \geqslant s^{J-M_{n-1}} = P[S_n|J, M_n] \tag{4.17}$$

with probability one. Hence,

$$P[S_{n+1}] \geqslant P[S_n], \; n = 1, 2, \ldots . \tag{4.18}$$

This result is intuitively clear. Each particle crossing a field in a certain path has higher survival probability than those of the particles crossing previously at the same path.

We remark here that if the Poisson approximation of the distribution of J is applied we obtain

$$P[S_1] = E\{s^J\} = \exp\{-\lambda(1-s)\}.$$

For all $n \geqq 2$ we obtain $P[S_n]$ by averaging $P[S_n|J=j]$ with respect to the Poisson probabilities $p(j|\lambda) = e^{-\lambda}\lambda^j/j!$, $j = 0, 1, \ldots$. In the numerical computations we sum over $j = 0, \ldots, JT$. If λ is small we fix a proper small value. For large λ we set $JT = INT(\lambda + 3\sqrt{\lambda}) + 1$, where $INT(\lambda)$ is the integer part of λ. Moreover, in the Poisson case we can treat the problem of defective absorption points (duds) very simply. If there are J points in a crossing path and the probability of "dud" is p_{dud} then the number of active points in the path, J^* has a binomial distribution with parameters J and $P = 1-p_{dud}$. Since J has a Poisson distribution with mean λ it is well known that J^* has (marginal) Poisson distribution with mean $\lambda^* = \lambda(1-p_{dud})$. Thus, we have to determine the expectations of the conditional probabilities $P(S_n|J)$ with respect to the Poisson distribution with mean λ^*.

Finally, the expected number of particles that survive crossing in a column of n can be obtained by adding their respective individual probabilities. Indeed, if we denote by I_i a random variable which assumes the value 1 if the i-th particles in a column survives and the value zero otherwise, then $P[S_i] = E\{I_i\}$, $i = 1, 2, \ldots, n$. Moreover, the total number of objects among n attempting to cross the field in a column which survive is $X_n = \sum_{i=1}^{n} I_i$. Accordingly, $E\{X_n\} = \sum_{i=1}^{n} P[S_i]$. In the following section we derive the distribution of X_n. However, for the determination of the expectation $E\{X_n\}$ that distribution is not needed. In the following table we provide numerical results obtained according to the recursive formulae developed here.

TABLE 1. *Survival Probabilities of Particles Crossing in a Column*

	Case I		Case II		Case III	
n	$P(N_n)$	$E\{X_n\}$	$P(S_n)$	$E\{X_n\}$	$P(S_n)$	$E\{X_n\}$
1	0.01850	0.0185	0.13601	0.1360	0.16843	0.1684
2	0.05306	0.0716	0.29144	0.4275	0.24209	0.4105
3	0.12852	0.2001	0.50502	0.9325	0.38551	0.7960
4	0.25060	0.4507	0.70332	1.6358	0.53058	1.3266
5	0.40656	0.8572	0.84330	2.4791	0.64798	1.9746
6	0.56973	1.4270	0.92188	3.4010	0.72686	2.7014
7	0.71347	2.1404	0.95770	4.3587	0.77236	3.4738
8	0.82211	2.9625	0.97101	5.3297	0.79553	4.2693
9	0.89328	3.8558	0.97504	6.3047	0.80623	5.0756
10	0.93381	4.7896	0.97606	7.2808	0.81080	5.8864

The parameters of the field are:

Case	λ	p_{det}	p_{act}	p_k
I	10	.25	.70	.80
II	5	.25	.70	.80
III	10	.25	.50	.50

5. THE DISTRIBUTION OF THE NUMBER OF SURVIVORS

In the present section we derive recursive formulae for the determination of the probability distribution of the (total) number of particles successfully crossing in a given path. Obviously $P[X_1 = 1|J] = s^J$. The joint conditional probability distribution of X_1 and M_1, given J, is

$$P[X_1 = 1,\ M_1 = r|J] = \begin{cases} w^J & , \text{ if } r = 0 \\ \binom{J}{r} w^{J-r}(s-w)^r & , \text{ if } r = 1, \ldots, J-r \\ (s-w)^J & , \text{ if } r = J \quad . \end{cases} \tag{5.1}$$

The derivation of (5.1) follows similar arguments to those of (4.16). Moreover,

$$P[X_1 = 0,\ M_1 = r|J] = P[M_1 = r|J] - P[X_1 = 1,\ M_1 = r|J]. \tag{5.2}$$

The joint conditional probability distribution of (X_n, M_n), given J, is determined recursively for $n = 2, 3, \ldots$ according to the formula:

$$\begin{aligned} &P[X_n = i,\ M_n = r|J] \\ &= \sum_{k=0}^{r} \Big\{ P[X_{n-1} = i-1,\ M_{n-1} = k|J]\ P[X_1 = 1,\ M_1 = r-k|J-k] \\ &\quad + P[X_{n-1} = i,\ M_{n-1} = k|J]\ P[X_1 = 0,\ M_1 = r-k|J-k] \Big\}; \end{aligned} \tag{5.3}$$

for each $i = 0, \ldots, n$; $r = 0, \ldots, J$; $J = 0, 1, \ldots$. From this joint conditional probability distribution of (X_n, M_n) given J we obtain the conditional probability distribution of X_n given J. Indeed,

$$(5.4) \qquad P[X_n = i|J] = \sum_{r=0}^{J} P[x_n = i, \ M_n = r|J].$$

Finally, according to the Poisson approximation

$$(5.5) \qquad P_\lambda[X_n = i] = \sum_{j=0}^{\infty} p(j|\lambda) \ P[X_n = i|J = j].$$

In the following table we provide some numberical results obtained by the above formulae.

TABLE 2. *The Probability Distribution of the Number of Survivors.* $n=1(1)10$, $\lambda = 2$, $p_{dud} = .05$, $p_{\text{det}} = .25$, $p_{act} = .7$ and $p_k = .8$.

n \ X_n	0	1	2	3	4	5	6	7	8	9	10
1	0.300	0.700	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.066	0.322	0.612	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.011	0.083	0.323	0.583	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.001	0.015	0.089	0.323	0.572	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.002	0.017	0.091	0.322	0.568	0.000	0.000	0.000	0.000	0.000
6	0.000	0.000	0.002	0.017	0.091	0.322	0.568	0.000	0.000	0.000	0.000
7	0.000	0.000	0.000	0.002	0.017	0.091	0.322	0.568	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.002	0.017	0.091	0.322	0.568	0.000	0.000
9	0.000	0.000	0.000	0.000	0.000	0.002	0.017	0.091	0.322	0.568	0.000
10	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.017	0.091	0.322	0.568

It is interesting to note that when λ is as small as in the above example the mode and the median of the survival distributions, for each $n=1, \ldots, 10$ is at $X_n = n$. Moreover, for all n greater than a certain number which depends on parameters of the example (here $n_0 = 4$) the distributions are related to each other by simple displacement.

6. EXAMPLE

In the present section we combine the procedures previously developed to determine the expected number of particles that survive, under the following field structure.

The field size is 200m × 200m. There are nine clusters centered at $\xi = -50, 0, 50$m and $\eta = -50, 0, 50$m. Each cluster consists of $N = 50$ absorption points, distributed according to a bivariate normal distribution with $\sigma_x = 10$m, $\sigma_y = 5$m and angle of orientation $\theta = 30°$. We compute the expected number of survivors in columns of $n = 10$ particles breaching in twenty paths of width $DL = 1$m spread along the field 10 meters apart. In the following table we provide the value of λ^* for each path and the expected number of survivors there. We also give the average number over the 20 paths. This average estimates the expected number of survivors if the location of the breaching path is chosen at random.

TABLE 3. *Expected Number of Survivors in Columns of* $n=10$, $p_{\text{det}} = .4$, $p_{act} = .95$, $p_k = .7$, $p_{dud} = .2$, $n=50$.

B	λ^*	$E\{X_n\}$
-95.	0.000	10.000
-85.	0.003	9.999
-75.	0.114	9.953
-65.	1.331	9.452
-55.	4.552	8.127
-45.	4.552	8.127
-35.	1.334	9.451
-25.	0.228	9.906
-15.	1.334	9.451
-5.	4.552	8.127
5.	4.552	8.127
15.	1.334	9.451
25.	0.228	9.906
35.	1.334	9.451
45.	4.552	8.127
55.	4.552	8.127
65.	1.331	9.452
75.	0.114	9.953
85.	0.003	9.999
95.	0.000	10.000
Average		9.259

Notice that $\lambda^* = \lambda(1-p_{dud})$. We see that the smallest number of expected survivors is around the aim point-rows at -50, 0 and 50. Due to the symmetric centering of the nine clusters around the center of the field the expected number of survivors, $E\{X_n\}$ is a symmetric function of the distance of the breaching point, B, from the middle of the field.

7. POSSIBLE GENERALIZATIONS

The model and algorithm discussed in the present paper can be easily generalized and extended to cover the following cases:

(i) Delivery errors: the centers of the clusters (ξ_i, η_i) are randomly distributed around their respective aim points.

(ii) The detection probabilities may change from one particle to another in a given column.

(iii) The detection probability of each particle may increase as the number of encounters of that particle grows.

(iv) Particles which move through the breaching paths in jumps (the personnel case).

Another interesting problem area concerns questions of optimal choice of the breaching path, when the information on the field structure is incomplete, the statistical analysis of the field structure (cluster analysis) from observations on the results of encounters, and similar problems of pattern recognition.

BIBLIOGRAPHY

[1] Briggs, B. D. "A General Computer Program for Use in Determining Track Width Plow-Minefield Criteria (U)," Proceedings of the Twelfth U.S. Army Operations Research Symposium, October 1973.

[2] Cudney, D. E. and D. O. Fraser, "Minefield Simulation (MINSIM I) Computer Model," Air Force Armament Laboratory, Technical Report AFATL-TR-128, Vol. I, II, Dec. 1971.

[3] Dean, K. J. and J. A. Cristians, "Battlefield Related Evaluation and Analysis of Countermine Hardware (Breach) (U)," Proceedings of the Twelfth U. S. Army Operations Research Symposium, October 1973.

[4] Heaps, W. E. and W. L. Warfield, "Minefield Effectiveness Models and Computer Programs for Personnel, Vehicular and Tank Targets (U)," AMSAA Technical Memo 109, June 1971. AMSAA, Aberdeen Proving Ground, Aberdeen, Maryland.

[5] Parsons, J. A. "Survival Probabilities Associated with Crossing Fields Containing Absorption Points," Naval Research Logistics Quarterly, *20*:41-51 (1973).

[6] Zacks, S. and D. Goldfarb, "Survival Probabilities in Crossing a Field Containing Absorption Points," Naval Research Logistics Quarterly, *13*:35-48 (1966).

[7] Zacks, S. "Bayes Sequential Strategies for Crossing a Field Containing Absorption Points," Naval Research Logistics Quarterly, *14*:329-43 (1967).

THE ASYMPTOTIC DISTRIBUTION OF ORDER STATISTICS*

Lionel Weiss

Cornell University
Ithaca, New York

ABSTRACT

For each n, $X_1(n)$, $X_2(n)$, . . . , $X_n(n)$ are IID, with common pdf $f_n(x)$. $Y_1(n) < \ldots < Y_n(n)$ are the ordered values of $X_1(n)$, . . . , $X_n(n)$. K_n is a positive integer, with $\lim_{n\to\infty} K_n = \infty$. Under certain conditions on K_n and $f_n(x)$, it was shown in an earlier paper that the joint distribution of a special set of $K_n + 1$ of the variables $Y_1(n)$, . . . , $Y_n(n)$ can be assumed to be normal for all asymptotic probability calculations. In another paper, it was shown that if $f_n(x)$ approaches the pdf which is uniform over (0, 1) at a certain rate as n increases, then the conditional distribution of the order statistics not in the special set can be assumed to be uniform for all asymptotic probability calculations. The present paper shows that even if $f_n(x)$ does not approach the uniform distribution as n increases, the distribution of the order statistics contained between order statistics in the special set can be assumed to be the distribution of a quadratic function of uniform random variables, for all asymptotic probability calculations. Applications to statistical inference are given.

1. NOTATION AND ASSUMPTIONS

For each positive integer n, the continuous random variables $X_1(n)$, . . . , $X_n(n)$ are IID, with $Y_1(n) < \ldots < Y_n(n)$ denoting their ordered values. For typographical simplicity, from now on we write $X_i(n)$, $Y_i(n)$ as X_i, Y_i respectively.

$f_n(x)$ and $F_n(x)$ denote respectively the common pdf and cdf for X_i. We assume that for all n, all x, and for $r = 1, 2, 3, 4$, $\frac{d^r}{dx^r} f_n(x)$ exists, and $\left|\frac{d^r}{dx^r} f_n(x)\right| < \Delta_1 < \infty$. We also assume that $f_n(x) < \Delta_2 < \infty$ for all n and all x.

For each n, we choose values p_n, q_n, and L_n to satisfy: $0 < p_n < q_n < 1$; L_n is a positive integer such that $\lim_{n\to\infty} \frac{L_n}{n^{2/3+\delta}} = 1$ for some fixed δ in the open interval $\left(0, \frac{1}{12}\right)$; np_n and nq_n are integers; $\frac{n(q_n - p_n)}{L_n} \equiv K_n$, say, is an integer; $\lim_{n\to\infty} p_n = 0$; $\lim_{n\to\infty} q_n = 1$; $\lim_{n\to\infty} np_n = \infty$; $\lim_{n\to\infty} n(1 - q_n) = \infty$; $\frac{f_n(F_n^{-1}(p_n))}{f_n(x)} > \Delta_3 > 0$ for all $x < F_n^{-1}(p_n)$; and $\frac{f_n(F_n^{-1}(q_n))}{f_n(x)} > \Delta_4 > 0$ for all $x > F_n^{-1}(q_n)$.

*Research supported by NSF Grant No. MCS76-06340.

Define b_n as $\inf_x \{f_n(x): F_n^{-1}(p_n) \leqslant x \leqslant F_n^{-1}(q_n)\}$. We assume that $n^{-\gamma} \leqslant b_n \leqslant 1$ for a fixed value γ in the open interval $\left(0, \min \left\{\frac{3\delta}{4}, \frac{1}{24} - \frac{\delta}{2}\right\}\right)$.

For $j = 1, \ldots, K_n + 1$, denote $F_n^{-1}\left(\frac{np_n + (j-1)L_n}{n}\right)$ by $T(j, n)$, denote $f_n(T(j, n))$ by $t(j, n)$, and denote $\sqrt{n}\; t(j, n)\; (Y_{np_n+(j-1)L_n} - T(j, n))$ by $Z_j(n)$. For $j = 1, \ldots, K_n$, denote $\frac{1}{2}\; (Y_{np_n+(j-1)L_n} + Y_{np_n+jL_n})$ by $\overline{Y}_j(n)$, and denote $(Y_{np_n+jL_n} - Y_{np_n+(j-1)L_n})$ by $D_j(n)$.

Let $W'(1, j, n), \ldots, W'(L_n - 1, j, n)$ denote the $L_n - 1$ values among $\{X_1, \ldots, X_n\}$ which fall in the open interval $\left(\overline{Y}_j(n) - \frac{D_j(n)}{2}, \overline{Y}_j(n) + \frac{D_j(n)}{2}\right)$, written in random order (that is, the same order in which $X_1, \ldots, X_n$ are written). For $i = 1, \ldots, L_n - 1$ and $j = 1, \ldots, K_n$, define $W(i, j, n)$ as $\frac{W'(i, j, n) - \overline{Y}_j(n)}{D_j(n)}$.

Let $g_{0,n}(z_1, \ldots, z_{K_n+1})$ denote the joint marginal pdf for $Z_1(n), \ldots, Z_{K_n+1}(n)$. $g_{0,n}$ is easily calculated, and is given explicitly in [3]. Let $g_{j,n}(w(1, j), \ldots, w(L_n - 1, j) | Z(n))$ denote the joint conditional pdf for $W(1, j, n), \ldots, W(L_n - 1, j, n)$, given $Z_1(n), \ldots, Z_{K_n+1}(n)$. (Note that being given $Z_1(n), \ldots, Z_{K_n+1}(n)$ is equivalent to being given $\overline{Y}_1(n), \ldots, \overline{Y}_{K_n}(n), D_1(n), \ldots, D_{K_n}(n)$.)

$$g_{j,n}(w(1, j), \ldots, w(L_n - 1, j) | Z(n)) =$$
$$\prod_{i=1}^{L_n - 1} \left\{ \frac{D_j(n) f_n(\overline{Y}_j(n) + D_j(n) w(i, j))}{F_n\left(\overline{Y}_j(n) + \frac{D_j(n)}{2}\right) - F_n\left(\overline{Y}_j(n) - \frac{D_j(n)}{2}\right)} \right\}$$

if $-\frac{1}{2} < w(i, j) < \frac{1}{2}$ for $i = 1, \ldots, L_n - 1$; $g_{j,n} = 0$ otherwise. Thus the joint pdf for the $K_n L_n + 1$ random variables $\{Z_1(n), \ldots, Z_{K_n+1}(n); W(i, j, n)$ for $i = 1, \ldots, L_n - 1$ and $j = 1, \ldots, K_n\}$ is $g_{0,n} \prod_{j=1}^{K_n} g_{j,n}$, which we denote by g_n.

Now we construct an "artificial" joint pdf for these $K_n L_n + 1$ random variables, as follows. The joint marginal pdf for $Z_1(n), \ldots, Z_{K_n+1}(n)$ is the joint normal pdf $h_{0,n}(z_1, \ldots, z_{K_n+1})$ defined as

$$\frac{C_n}{(\sqrt{2\pi})^{K_n+1}} \exp\left\{ -\frac{n(L_n - 1)}{2L_n^2} \left[\frac{L_n z_1^2}{np_n} + \frac{L_n z_{K_n+1}^2}{n(1 - q_n)} + \sum_{j=1}^{K_n} (z_{j+1} - z_j)^2 \right] \right\}$$

where the value C_n is given explicitly in [3]. Under the artificial distribution, the $K_n(L_n - 1)$ random variables $\{W(i, j, n)\}$ are mutually independent, and are independent of $Z_1(n), \ldots, Z_{K_n+1}(n)$, with distributions constructed as follows. Let $U(i, j)$ $(i = 1, \ldots, L_n - 1;\ j = 1, \ldots, K_n)$ be independent random variables, each uniformly distributed over

(0, 1), and independent of $(Z_1(n), \ldots, Z_{K_n+1}(n))$. Denote $F_n^{-1}\left(\dfrac{np_n + \left(j - \frac{1}{2}\right)L_n}{n}\right)$ by $\alpha_j(n)$, and $\dfrac{L_n}{2n}\,\dfrac{f_n'(\alpha_j(n))}{f_n^2(\alpha_j(n))}$ by $\gamma_j(n)$, for $j = 1, \ldots, K_n$. Then the distribution of $W(i, j, n)$ is to be the distribution of $-\frac{1}{2} + (1 + \gamma_j(n))\, U(i, j) - \gamma_j(n)\, U^2(i, j)$. Assuming that n is large enough so that $|\gamma_j(n)| < 1$ for $j = 1, \ldots, K_n$, the pdf for $W(i, j, n)$ is then $[1 + \gamma_j^2(n) - 4\gamma_j(n)\, w(i, j)]^{-1/2}$ for $-\frac{1}{2} < w(i, j) < \frac{1}{2}$, and is zero otherwise. Denote the pdf by $h_{j,n}(w(i, j))$. Then the artificial joint pdf for all $K_nL_n + 1$ random variables is $h_{0,n}(z_1, \ldots, z_{K_n+1}) \prod_{j=1}^{K_n} \prod_{i=1}^{L_n-1} h_{j,n}\, w(i, j)$, which we denote by h_n.

If S_n is an arbitrary measurable region in $(K_nL_n + 1)$-dimensional space, let $P_{g_n}(S_n)$, $P_{h_n}(S_n)$ denote respectively the probabilities assigned to S_n by the joint densities g_n, h_n. The next section is devoted to proving the following Theorem:

$$\lim_{n\to\infty} \sup_{S_n} |P_{g_n}(S_n) - P_{h_n}(S_n)| = 0.$$

2. PROOF OF THE THEOREM

Throughout this section, it will always be assumed that the joint distribution of the $K_nL_n + 1$ random variables is given by the density h_n. As stated in [3], the theorem will be proved if we show that

$$\log \frac{g_n(Z_1(n), \ldots, Z_{K_n+1}(n), W(1, 1, n), \ldots, W(L_n - 1, K_n, n))}{h_n(Z_1(n), \ldots, Z_{K_n+1}(n), W(1, 1, n), \ldots, W(L_n - 1, K_n, n))} \tag{2.1}$$

converges stochastically to zero as n increases. Let $R_n(i, j, W(i, j, n))$ denote

$$\log\left[\left\{\frac{D_j(n) f_n(\bar{Y}_j(n) + D_j(n)\, W(i, j, n))}{F_n\left(\bar{Y}_j(n) + \dfrac{D_j(n)}{2}\right) - F_n\left(\bar{Y}_j(n) - \dfrac{D_j(n)}{2}\right)}\right\}\sqrt{1 + \gamma_j^2(n) - 4\gamma_j(n)\, W(i, j, n)}\right].$$

Then (2.1) can be written as the sum of the two expressions

$$\log\left\{\frac{g_{0,n}(Z_1(n), \ldots, Z_{K_n+1}(n))}{h_{0,n}(Z_1(n), \ldots, Z_{K_n+1}(n))}\right\} \tag{2.2}$$

$$\sum_{j=1}^{K_n} \sum_{i=1}^{L_n-1} R_n(i, j, W(i, j, n)). \tag{2.3}$$

The fact that the expression (2.2) converges stochastically to zero as n increases is the theorem of [3]; it is easily verified that the assumptions of [3] are implied by our present assumptions. Thus we only have to verify that the expression (2.3) converges stochastically to zero as n increases.

Let β_n denote $\inf\{f_n(x)\colon Y_{np_n} \leqslant x \leqslant Y_{nq_n}\}$. We are going to compute the conditional mean and the conditional variance of the expression (2.3) given $\{Z_1(n), \ldots,$

$Z_{K_n+1}(n)\} \equiv Z(n)$, say, under the artificial distribution. Denote the conditional mean and the conditional variance, given $Z(n)$, of the expression (2.3) by $m_n(Z(n))$ and $v_n(Z(n))$ respectively. Computing $E\{R_n(i,j,W(i,j,n)|Z(n)\}$ and $E\{R_n^2(i, j, W(i, j, n)|Z(n)\}$ by the formulas $\int_{-1/2}^{1/2} R_n(i, j, w)h_{j,n}(w)dw$ and $\int_{-1/2}^{1/2} R_n^2(i, j, w)h_{j,n}(w)dw$, tedious but routine calculations show that these two integrals can be written respectively as

$$-\frac{1}{24}\left[\frac{f_n'(\bar{Y}_j(n))}{f_n(\bar{Y}_j(n))} D_j(n) - 2\gamma_j(n)\right]^2 + \sum_{i=0}^{4} V_{i,j}(n)\gamma_j^i(n)\left[\frac{D_j(n)}{\beta_n}\right]^{4-i},$$

and

$$\frac{1}{12}\left[\frac{f_n'(\bar{Y}_j(n))}{f_n(\bar{Y}_j(n))} D_j(n) - 2\gamma_j(n)\right]^2 + \sum_{i=0}^{4} V_{i,j}^*(n)\gamma_j^i(n)\left[\frac{D_j(n)}{\beta_n}\right]^{4-i}$$

where there is a finite constant V such that $|V_{i,j}(n)| < V$ and $|V_{i,j}^*(n)| < V$ for $i = 0, 1, 2, 3, 4$, and for all j and n. Noting that, under the artificial distribution, the joint conditional distribution of $\{R_n(i, j, W(i, j): i = 1, \ldots, L_n - 1; j = 1, \ldots, K_n\}$, given $Z(n)$, is that of mutually independent random variables, we find

$$m_n(z(n)) = -\frac{(L_n - 1)}{24}\sum_{j=1}^{K_n}\left[\frac{f_n'(\bar{Y}_j(n))}{f_n(\bar{Y}_j(n))} D_j(n) - 2\gamma_j(n)\right]^2$$

$$+ (L_n - 1)\sum_{j=1}^{K_n}\sum_{i=0}^{4} V_{i,j}(n)\gamma_j^i(n)\left[\frac{D_j(n)}{\beta_n}\right]^{4-i} \text{ and}$$

$$v_n(Z(n)) = \frac{(L_n - 1)}{12}\sum_{j=1}^{K_n}\left[\frac{f_n'(\bar{Y}_j(n))}{f_n(\bar{Y}_j(n))} D_j(n) - 2\gamma_j(n)\right]^2$$

$$+ (L_n - 1)\sum_{j=1}^{K_n}\sum_{i=0}^{4} \bar{V}_{i,j}(n)\gamma_j^i(n)\left[\frac{D_j(n)}{\beta_n}\right]^{4-i},$$

where there exists a finite constant $\bar{V}$ such that $|\bar{V}_{i,j}(n)| < \bar{V}$ for $i = 0, 1, 2, 3, 4$ and all j and n.

The rest of this section is devoted to showing that $m_n(Z(n))$ and $v_n(Z(n))$ both converge stochastically to zero as n increases, which clearly implies that the expression (2.3) converges stochastically to zero as n increases, thus completing the proof of the theorem.

First we investigate β_n. If $Y_{np_n} < F_n^{-1}(p_n)$, let y be any value in $[Y_{np_n}, F_n^{-1}(p_n)]$, so $y = F_n^{-1}(p_n) + \dfrac{\theta Z_1(n)}{\sqrt{n} f_n(F_n^{-1}(p_n))}$, with $-1 \leqq \theta \leqq 0$. Then we can write $f_n(y) =$

$$f_n(F_n^{-1}(p_n)) + \frac{f_n'(y^*)\theta\sqrt{p_n}}{\sqrt{n}\ f_n(F_n^{-1}(p_n))}\left[\frac{Z_1(n)}{\sqrt{p_n}}\right].$$

It was shown in [3] that $\dfrac{Z_1(n)}{\sqrt{p_n}}$ has asymptotically a standard normal distribution, and since p_n approaches zero as n increases, it follows that with probability approaching one as n increases, $f_n(y) \geqq b_n - \dfrac{1}{\sqrt{n}\, b_n}$. Similarly, if $Y_{nq_n} > F_n^{-1}(q_n)$, we can show that for any value y in $[F_n^{-1}(q_n), Y_{nq_n}]$, $f_n(y) \geqq b_n - \dfrac{1}{\sqrt{n}\, b_n}$ with probability approaching one as n increases. Since $b_n \geqq n^{-\gamma}$, it follows that with probability approaching one as n increases, $\beta_n \geqq \dfrac{1}{2} n^{-\gamma}$.

Let $\Phi(x)$ denote the standard normal cdf. In [3] it was shown that Variance $(Z_j(n) - Z_{j-1}(n)) = \frac{L_n}{n}(1 + \delta_n'')$, where δ_n'' approaches zero as n increases, and that Variance $(Z_j(n)) \leqq \overline{\sigma}^2 < \infty$, for $j = 1, \ldots, K_n + 1$. From these facts, and from the inequalities $P(E_1 \cap \ldots \cap E_m) \geqq 1 - \sum_{i=1}^{m} [1 - P(E_i)]$ for any events $E_1, \ldots, E_m$, and $\Phi(x) \leqq 1 - \frac{1}{\sqrt{2\pi x}} e^{-x^2/2}$ for all $x > 0$, it follows easily that if c_1, c_2 are any positive values,

$$P[\max_{j=1,\ldots,K_n} |Z_j(n) - Z_{j-1}(n)| < n^{-\frac{1}{6}+\frac{\delta}{2}+c_1}]$$

and

$$P[\max_{j=1,\ldots,K_n+1} |Z_j(n)| < n^{c_2}]$$

both approach one as n increases. We take a value c_1 in the open interval $\left(0,\ \gamma + \frac{3\delta}{2}\right)$ and a value c_2 in the open interval $(0, \delta - \gamma)$ for use below.

From the definitions above, we have

$$\overline{Y}_j(n) = \frac{1}{2}(T(j+1, n) + T(j, n)) + \frac{1}{2\sqrt{n}}\left\{\frac{Z_{j+1}(n)}{t(j+1, n)} + \frac{Z_j(n)}{t(j, n)}\right\},$$

$$D_j(n) = T(j+1, n) - T(j, n) + \frac{1}{\sqrt{n}}\left\{\frac{Z_{j+1}(n)}{t(j+1, n)} - \frac{Z_j(n)}{t(j, n)}\right\}.$$

Simple expansions give

$$T(j+1, n) = \alpha_j(n) + \frac{L_n}{2nf_n(\alpha_j(n))} + \hat{V}_{1,j}(n)\left(\frac{L_n}{nb_n}\right)^2,$$

$$T(j, n) = \alpha_j(n) - \frac{L_n}{2nf_n(\alpha_j(n))} + \hat{V}_{2,j}(n)\left(\frac{L_n}{nb_n}\right)^2,$$

$$t(j+1, n) = f_n(\alpha_j(n)) + \frac{L_n}{2n}\,\frac{f_n'\alpha_j(n))}{f_n(\alpha_j(n))} + \hat{V}_{3,j}(n)\left(\frac{L_n}{nb_n}\right)^2,$$

$$t(j, n) = f_n(\alpha_j(n)) - \frac{L_n}{2n}\,\frac{f_n'(\alpha_j(n))}{f_n(\alpha_j(n))} + \hat{V}_{4,j}(n)\left(\frac{L_n}{nb_n}\right)^2,$$

where $|\hat{V}_{i,j}(n)| < \hat{V} < \infty$ for $i = 1, 2, 3, 4$, all n and j, and some constant $\hat{V}$. Using these facts, and the results above on the orders of magnitude of $\max_j |Z_j(n) - Z_{j-1}(n)|$, $\max_j |Z_j(n)|$, and β_n, we find that

$$\max_{j=1,\ldots,K_n}\left|D_j(n) - \frac{L_n}{nf_n(\alpha_j(n))}\right| = O_p(n^{-2/3+2\gamma+2\delta}),$$

$$\max_{j=1,\ldots,K_n} |\overline{Y}_j(n) - \alpha_j(n)| = O_p(n^{c_2-1/2+\gamma}),$$

and

$$\max_{j=1,\ldots,K_n}\left|\frac{f_n'(\overline{Y}_j(n))}{f_n(\overline{Y}_j(n))} - \frac{f_n'(\alpha_j(n))}{f_n(\alpha_j(n))}\right| = O_p(n^{c_2-1/2+3\gamma}),$$

where O_p has the usual meaning. Using these order relationships in the expressions for $m_n(Z(n))$ and $v_n(Z(n))$ given above, we find that both these expressions converge stochastically to zero as n increases, completing the proof.

3. EARLIER RESULTS, AND EXTENSIONS

In [4], it was shown that if $f_n(x)$ assigned probability one to the unit interval, and $\max_{0 \leq x \leq 1} |f_n(x) - 1|$ approached zero rapidly enough, in the artificial distribution we can set $\gamma_j(n) = 0$ for all n and all j.

In the present case, several generalizations are possible. A few will now be sketched.

If $f_n(x)$ assigns probability one to a bounded interval, and is bounded away from zero over the interval, we can treat the order statistics below Y_{np_n} and above Y_{nq_n} in the same manner as we treated the order statistics falling between $Y_{np_n+(j-1)L_n}$ and $Y_{np_n+jL_n}$. This was done in [4] for the special case where $f_n(x)$ approaches the uniform density.

Several of the assumptions made above are convenient but not necessary. For example, it is not necessary that b_n be less than or equal to one. It was assumed that $b_n \leqslant 1$ in [3] merely to avoid having to modify the argument to take account of the two separate cases, $b_n \leqslant 1$ and $b_n > 1$. Similarly, it is not necessary that p_n approach zero as n increases, or that q_n approach one as n increases.

Under certain conditions, it is possible to take L_n asymptotically equivalent to $n^{1/2+\delta}$, as in [1].

4. APPLICATIONS

In the applications, the actual joint distribution is g_n, but the artificial distribution h_n is used for asymptotic probability calculations, because it is simpler and more convenient.

As a first application, suppose $f_n(x) = \frac{1}{\sigma} \bar{f}\left(\frac{x-\mu}{\sigma}\right)$, where μ and σ are unknown parameters (with $\sigma > 0$), and $\bar{f}$ is a known function. It is easily verified that in this case, $\gamma_j(n)$ does not depend on μ or σ, which means that asymptotically, if we are given the values $\{Y_{np_n}, Y_{np_n+L_n}, \ldots, Y_{nq_n}\}$, then the other order statistics falling between Y_{np_n} and Y_{nq_n} contain no additional information about μ and σ. In some cases, such as when $\bar{f}$ is a normal pdf, the order statistics below Y_{np_n} and above Y_{nq_n} asymptotically contain no information about μ and σ if p_n approaches zero and q_n approaches one at arbitrary rates as n increases. In such cases, asymptotically efficient estimators of μ and σ can be based on $\{Y_{np_n}, Y_{np_n+L_n}, \ldots, Y_{nq_n}\}$ alone. Such estimators can usually be given as linear functions of $\{Y_{np_n}, Y_{np_n+L_n}, \ldots, Y_{nq_n}\}$, as in [3].

A second application is an extension of the first. Suppose $f_n(x) = \frac{1}{\sigma} \bar{f}\left(\frac{x-\mu}{\sigma}; \theta_1, \ldots, \theta_m\right)$, where $\mu, \sigma, \theta_1, \ldots, \theta_m$ are unknown parameters (with $\sigma > 0$ and possibly some restrictions on $\theta_1, \ldots, \theta_m$), and $\bar{f}$ is a known function. In this case, $\gamma_j(n)$ depends on $\theta_1, \ldots, \theta_m$, but not on μ or σ; we write it as $\gamma_j(\theta_1, \ldots, \theta_m; n)$. In some cases, it is possible to construct estimators of $\theta_1, \ldots, \theta_m$, based on $\{Y_{np_n}, Y_{np_n+L_n}, \ldots, Y_{nq_n}\}$, with the following property. Denote the estimators by $\bar{\theta}_1(n), \ldots, \bar{\theta}_m(n)$, and denote $\gamma_j(\bar{\theta}_1(n), \ldots, \bar{\theta}_m(n); n)$

by $\bar{\gamma}_j(n)$. Finally, denote by $\bar{h}_{j,n}(w(i, j))$ the pdf which is equal to $[1 + \bar{\gamma}_j(n) - 4\bar{\gamma}_j(n) w(i, j)]^{-1/2}$ if $-\frac{1}{2} < w(i, j) < \frac{1}{2}$, and is zero otherwise. The property desired is that

$$\log \prod_{j=1}^{K_n} \prod_{i=1}^{L_n-1} \left\{ \frac{\bar{h}_{j,n}(W(i, j))}{h_{j,n}(W(i, j))} \right\}$$

converges stochastically to zero as n increases, assuming that $W(i, j)$ are independent, with pdf for $W(i, j)$ given by $h_{j,n}(w(i, j))$. If this property holds, then asymptotically, if we are given the values $\{Y_{np_n}, Y_{np_n+L_n}, \ldots, Y_{nq_n}\}$, the other order statistics falling between Y_{np_n} and Y_{nq_n} contain no additional information about μ, σ, $\theta_1, \ldots, \theta_m$. This is so because if we are given $\{Y_{np_n}, Y_{np_n+L_n}, \ldots, Y_{nq_n}\}$, we can construct the pdfs $\bar{h}_{j,n}$, and then generate random variables $\bar{W}(i, j)$ whose joint pdf is $\prod_{j=1}^{K_n} \prod_{i=1}^{L_n-1} \bar{h}_{j,n}(\bar{w}(i, j))$. But by the property assumed, and the theorem above, we are just as well off asymptotically as if we knew all the order statistics between Y_{np_n} and Y_{nq_n}, since asymptotically all probabilities are the same for $\{\bar{W}(i, j)\}$ as for $\{W(i, j)\}$.

In particular, in many cases we can find estimators $\bar{\theta}_i(n)$ such that for $i = 1, \ldots, m$, $|\bar{\theta}_i(n) - \theta_i| = O_p\left(\frac{1}{\sqrt{n}}\right)$, and such that this implies that

$$\max_{j=1,\ldots,K_n} \left| \frac{n}{L_n} \bar{\gamma}_j(n) - \frac{n}{L_n} \gamma_j(\theta_1, \ldots, \theta_m; n) \right| = O_p\left(\frac{1}{\sqrt{n}}\right).$$

In such cases, the estimators can be shown to have the desired property.

For the next two applications, we introduce the following notation: $\theta_j(n)$ is to denote $\frac{f_n'(\alpha_j(n))}{f_n^2(\alpha_j(n))}$. The observable random variables $W(1, j, n), \ldots, W(L_n - 1, j, n)$ are IID with common pdf

$$\left\{1 + \left(\frac{L_n}{2n}\right)^2 \theta_j^2(n) - 4\left(\frac{L_n}{n}\right) \theta_j(n) w\right\}^{-1/2}$$

for $-\frac{1}{2} < w < \frac{1}{2}$. Denote the observable quantity

$$\frac{12n}{L_n(L_n - 1)} \sum_{i=1}^{L_n-1} W(i, j, n)$$

by $\hat{\theta}_j(n)$. Using the representation of $W(i, j, n)$ as

$$-\frac{1}{2} + (1 + \frac{L_n}{2n} \theta_j(n)) U(i, j) - \frac{L_n}{2n} \theta_j(n) U^2(i, j),$$

it is easily shown that $E\{\hat{\theta}_j(n)\} = \theta_j(n)$, and

$$\text{Variance }\{\hat{\theta}_j(n)\} = \frac{12n^2}{L_n^2(L_n - 1)} \left\{1 - \frac{L_n}{2n} \theta_j(n) + \frac{1}{15}\left(\frac{L_n}{2n} \theta_j(n)\right)^2\right\},$$

which is asymptotically equivalent to $\frac{12n^2}{L_n^3}$. A simple calculation shows that this is asymptotically equivalent to the Cramér-Rao lower bound on the variance of an unbiased estimator of

$\theta_i(n)$ based on $\{W(1, j, n), \ldots, W(L_n - 1, j, n)\}$. By the central limit theorem, the asymptotic distribution of $\frac{L_n^{3/2}}{\sqrt{12n}} (\hat{\theta}_i(n) - \theta_i(n))$ is standard normal.

As a third application, suppose we are given a fixed value b in the open interval (0, 1), and we want to test the hypothesis that $\frac{f_n'(F_n^{-1}(b))}{f_n^2(F_n^{-1}(b))} = 0$. (This may be part of a test that F_n has a mode at $F_n^{-1}(b)$). Suppose we choose p_n, q_n, and L_n so that

$$\frac{np_n + \left(j - \frac{1}{2}\right)L_n}{n} = b$$

for some integer j between 1 and K_n, say $j = J$. Then

$$\frac{f_n'(F_n^{-1}(b))}{f_n^2(F_n^{-1}(b))} = \theta_J(n),$$

and a two-sided test of the hypothesis with asymptotic level of significance α is to accept if

$$|\hat{\theta}_i(n)| < \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \frac{\sqrt{12n}}{L_n^{3/2}}.$$

A fourth application is to the device called "hedging" in an earlier paper [2]. This is used to modify an estimator if the data suggest our model is inaccurate. Thus, suppose the problem is to estimate the unknown median of a distribution. If we think the distribution is normal (with unspecified mean and variance), an asympototically efficient estimator of the median is the sample mean, but a certain linear combination of $\{Y_{np_n}, Y_{np_n+L_n}, \ldots, Y_{nq_n}\}$, mentioned in the first application above, is also asymptotically efficient. Denote this linear combination by $Q(n)$. On the other hand, if the distribution is not normal, $Q(n)$ may be a very poor estimator of the population median. We now describe a hedging device for this problem.

Choose p_n, q_n, and L_n so that $np_n + jL_n = \left[\frac{n}{2}\right]$ for some j between 1 and K_n, so that the sample median is one of the variables $\{Y_{np_n}, Y_{np_n+L_n}, \ldots, Y_{nq_n}\}$. If $f_n(x)$ is actually a normal density, then $\theta_i(n)$ is equal to the particular value $\theta_j^*(n)$ defined as

$$\sqrt{2\pi}\, \Phi^{-1}\left(\frac{np_n + \left(j - \frac{1}{2}\right)L_n}{n}\right) \exp\left\{\frac{1}{2}\left[\Phi^{-1}\left(\frac{np_n + \left(j - \frac{1}{2}\right)L_n}{n}\right)\right]^2\right\}.$$

Let $S(n)$ denote the observable random variable

$$\sum_{j=1}^{K_n} \left[\frac{\hat{\theta}_j(n) - \theta_j^*(n)}{\sqrt{\frac{12n^2}{L_n^3}}}\right]^2,$$

and let $S^*(n)$ denote the nonrandom nonnegative quantity

$$\sum_{j=1}^{K_n} \left[\frac{\theta_j(n) - \theta_j^*(n)}{\sqrt{\frac{12n^2}{L_n^3}}}\right]^2.$$

Using the results above, it is easily shown that the asymptotic distribution of

$$\frac{S(n) - (K_n + S_n^*(n))}{\sqrt{2K_n + 4S^*(n)}}$$

is standard normal. Thus an observed $S(n)$ which is close to K_n is evidence that f_n is a normal density. The hedging device is to use $Q(n)$ as the estimator of the population median if $S(n) < K_n + C_n^*$, and use the sample median if $S(n) \geqq K_n + C_n^*$, for some nonrandom C_n^*. The choice of C_n^*, and the analysis of the properties of the overall method of estimation, can be carried out as in [2]. The analysis is particularly easy because the joint asymptotic distribution of $S(n)$, $Q(n)$, and the sample median is trivariate normal, with $S(n)$ independent of the two other random variables.

REFERENCES

[1] Reiss, R.D., "The Asymptotic Normality and Asymptotic Expansions for the Joint Distribution of Several Order Statistics," Colloquium Mathematics Society Janos Bolyai (1974).

[2] Weiss, L. "'Hedging' on Statistical Assumptions," Naval Research Logistics Quarterly (1961).

[3] Weiss, L'. "Statistical Procedures Based on a Gradually Increasing Number of Order Statistics," Communications in Statistics (1973).

[4] Weiss, L. "The Asymptotic Sufficiency of a Relatively Small Number of Order Statistics in Tests of Fit," Annals of Statistics (1974).

NONEXTREME POINT SOLUTION STRATEGIES FOR LINEAR PROGRAMS

Leon Cooper and Jeff Kennington

Department of Operations Research
and
Engineering Management
Southern Methodist University
Dallas, Texas

ABSTRACT

This exposition presents two algorithms for linear programs which allow a value change in more than one nonbasic variable at each iteration. The computational formulae are developed and errors which have appeared in the literature are noted. One algorithm is a multiple basis exchange procedure while the second is a feasible direction method. There remain many computational challenges in the area of linear programming and we hope that this investigation will encourage additional work in the directions indicated in this exposition.

1. INTRODUCTION

In 1947 George B. Dantzig developed the primal simplex algorithm for solving linear programs. Variations of this algorithm have been developed, such as the dual simplex method and the primal-dual method; but no other algorithm has ever seriously challenged this method for solving linear programs. Competing algorithms, such as Harold Kuhn's *Hungarian Algorithm* [4] and Delbert Fulkerson's *Out-of-Kilter Algorithm* [3] have been proposed for network programs, but generalizations of these procedures have been abandoned.

We find it curious that the literature contains so few papers concerning other algorithms for such an important class of problems. We assume that either (i) other ideas have been investigated, abandoned, and never reported, or (ii) the simplex method has proved so effective that other investigators felt no motivation to work in this area. In either case we believe that the issue should not be closed and offer two basic strategies that can be used in algorithmic development for this problem.

The two strategies presented in this paper were motivated by an attempt to make big jumps rather than "crawling along the edges" of the convex set. The first strategy involves multiple column exchanges in the basis rather than the single exchange used by the simplex algorithm, while the second strategy involves cutting across the middle of the feasible region. Algorithms based on each of these strategies are developed.

Our purpose in presenting the solution strategies and algorithms embodied in this paper is to open up, for consideration, discussion and algorithmic development, the idea of nonextreme point solution strategies for linear programs. The development process for computationally effective algorithms for large-scale problems as the history of the "simplex algorithm" itself reveals, is a lengthy process to which many people ultimately contribute. In the course of investigation, new ideas suggest themselves and extensive changes are made to algorithms, as well as computational strategies for implementing the algorithms. The algorithms presented in this paper should be viewed as points of departure. One of them is under active investigation at present. It is hoped that others may also wish to conduct investigations in nonextreme point solution strategies.

2. MULTIPLE BASIS ENTRY

The linear programming problem is:

$$\text{(1)} \qquad \left.\begin{array}{rl} & \max z = \bar{c}'\bar{x} \\ \text{subject} & A\bar{x} = \bar{b} \\ & \bar{x} \geqslant \bar{0}, \end{array}\right\}$$

where A is an $m \times n$ matrix, $\bar{c}$ and $\bar{x}$ are n-vectors, and $\bar{b}$ is an m-vector. In (1) we assume that $r(A) = m$ and $m < n$.

A set of m linearly independent columns of A constitute a basis matrix B and so the constraints $Ax = b$ can be rewritten:

$$\text{(2)} \qquad (B,N)\begin{bmatrix}\bar{x}_B \\ \bar{x}_N\end{bmatrix} = \bar{b}.$$

Since B has an inverse, we can solve (2) for $\bar{x}_B$, yielding:

$$\text{(3)} \qquad \bar{x}_B = B^{-1}\bar{b} - B^{-1}N\bar{x}_N.$$

Since $\bar{x}_N = \bar{0}$ for a basic solution, (3) reduces to

$$\text{(4)} \qquad \bar{x}_B = B^{-1}\bar{b}.$$

In addition, the matrix B is chosen so that $\bar{x}_B \geqslant \bar{0}$. In the primal simplex algorithm one chooses some column $\bar{a}_k$ of N to enter the basis and one column $\bar{b}_r$ of B to be removed in such a way that the value of the objective function $\hat{z}$, after the change of basis, is greater than z. It is well known that $\hat{z} > z$ if $\bar{a}_k$ is a vector for which:

$$\text{(5)} \qquad \frac{x_{Br}}{y_{rk}}(c_k - z_k) > 0, \quad x_{Br} > 0, \quad y_{rk} > 0,$$

where x_{Br} is the r^{th} component of $\bar{x}_B$, y_{rk} is the r^{th} component of $\bar{y}_k$, $\bar{y}_k = B^{-1}\bar{a}_k$, $z_k = \bar{c}_B'\bar{y}_k$ and c_k is the k^{th} component of $\bar{c}$. Hence the new basis $\hat{B}$ that is obtained from B, after this exchange of vectors is:

$$\hat{B} = [\bar{b}_1, \bar{b}_2, \ldots, \bar{b}_{r-1}, \bar{a}_k, \bar{b}_{r+1}, \ldots, \bar{b}_m].$$

The vector $\bar{b}_r$ to be removed from the basis, is chosen in such a way that the new basic solution, $\bar{x}_B = \hat{B}^{-1}\bar{b} \geqslant 0$, i.e., it is a basic feasible solution.

What we wish to consider in this section, is inserting a set of vectors into the basis and removing a corresponding number of vectors from the basis, in such a way that the new basic solution is feasible and the new objective function has a greater value than the previous objective function. In brief, we wish to insert a set of vectors, say S_k, and remove a set S_r of the current basis vectors. In order to do so, we need to know the following, which are the analogues of the conditions of the simplex algorithm:

1) Under what conditions is the new set of vectors a basis?
2) Under what conditions will the new objective function be increased?
3) Under what conditions is the new solution feasible?

2.1 The Existence of a Basis

First we shall examine the question of maintaining a basis in the multiple entry exchange of vectors. The following theorem states the condition for maintaining a basis.

THEOREM 1: Let $B = [\bar{b}_1, \bar{b}_2, \ldots, \bar{b}_m]$ be a basis for E^m. Then $\hat{B} = [\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_p, \bar{b}_{p+1}, \ldots, \bar{b}_m]$ is also a basis for E^m if and only if $P = B^{-1}\hat{B}$ is nonsingular.

PROOF: Given P nonsingular, we wish to prove that $\hat{B}$ is a basis. Suppose $\hat{B}$ is not a basis. Then $\det(P) = \det(B^{-1}\hat{B}) = \det(B^{-1})\det(\hat{B}) = 0$ which contradicts P nonsingular. Hence $\hat{B}$ is a basis.

Given that $\hat{B}$ is a basis, we wish to prove that P is nonsingular. Clearly, $\det(B^{-1}\hat{B}) = \det(B^{-1})\det(\hat{B}) \neq 0$ implies P nonsingular.

Theorem 1 provides a necessary and sufficient condition for a basis to exist if we replace several basis vectors at a time. It does not, however, insure that the basis will result in a feasible solution.

2.2 Updating Formulae

For our multiple basis entry technique, we propose to maintain only the current basis inverse and the current solution. In this section we develop the updating formulae required for maintaining these quantities if p vectors are exchanged in the basis.

Recall that the basic variables are given by

$$\bar{x}_B = B^{-1}\bar{b} - B^{-1}N\bar{x}_N.$$

We partition $\bar{x}_B$ and $\bar{x}_N$ as follows:

$$\bar{x}_B = \left(\begin{array}{c} \bar{x}_{B1} \\ \hline \bar{x}_{B2} \end{array}\right), \quad \bar{x}_N = \left(\begin{array}{c} \bar{x}_{N1} \\ \hline \bar{x}_{N2} \end{array}\right)$$

where $\bar{x}_{B1}$ will correspond to the p columns of the basis B, that are to be replaced by p columns of N that correspond to $\bar{x}_{N1}$. In other words, B and N are partitioned as:

$$B = \left[\begin{array}{c|c} p \times p & p \times (m-p) \\ \hdashline (m-p) \times p & (m-p) \times (m-p) \end{array}\right] = \left[\begin{array}{c|c} B_{11} & B_{12} \\ \hdashline B_{21} & B_{22} \end{array}\right]$$

and

$$N = \left|\begin{array}{c|c} p \times p & p \times (n-m-p) \\ \hline (m-p) \times p & (m-p) \times \\ & (n-m-p) \end{array}\right| = \left|\begin{array}{c|c} N_{11} & N_{12} \\ \hline N_{21} & N_{22} \end{array}\right|.$$

Since we are replacing the first p columns of B with the first p columns of N, it is clear that:

$$\hat{B} = \left|\begin{array}{c|c} N_{11} & B_{12} \\ \hline N_{21} & B_{22} \end{array}\right|$$

We also note that:

$$B^{-1}N = Y_N = \left|\begin{array}{c|c} Y_{11} & Y_{12} \\ \hline Y_{21} & Y_{22} \end{array}\right|.$$

Then from Theorem 1,

$$P = B^{-1}\hat{B} = \left|\begin{array}{c|c} Y_{11} & 0 \\ \hline Y_{21} & I \end{array}\right|,$$

must be nonsingular, if $\hat{B}$ is a basis. Therefore Y_{11} must be nonsingular, and

(7)
$$\hat{B}^{-1} = P^{-1}B^{-1},$$

where

(8)
$$P^{-1} = \left|\begin{array}{c|c} Y_{11}^{-1} & 0 \\ \hline -Y_{21}Y_{11}^{-1} & I \end{array}\right|.$$

We may now calculate $\hat{\bar{x}}_B$ from: $\hat{\bar{x}}_B = \hat{B}^{-1}\bar{b} = (BP)^{-1}\bar{b} = P^{-1}B^{-1}\bar{b} = P^{-1}\bar{x}_B$.

Hence

(9)
$$\hat{\bar{x}}_B = P^{-1}\bar{x}_B.$$

From (8) and (9) we obtain:

$$\hat{\bar{x}}_B = \left[\begin{array}{c|c} Y_{11}^{-1} & 0 \\ \hline -Y_{21}Y_{11}^{-1} & I \end{array}\right]\left[\begin{array}{c} \bar{x}_{B1} \\ \hline \bar{x}_{B2} \end{array}\right] = \left[\begin{array}{c} Y_{11}^{-1}\bar{x}_{B1} \\ \hline \bar{x}_{B2} - Y_{21}Y_{11}^{-1}\bar{x}_{B1} \end{array}\right].$$

Therefore we have:

(10)
$$\hat{\bar{x}}_{B1} = Y_{11}^{-1}\bar{x}_{B1}$$
$$\hat{\bar{x}}_{B2} = \bar{x}_{B2} - Y_{21}Y_{11}^{-1}\bar{x}_{B1}.$$

We now have formulae for maintaining the basis inverse, (7), and updating the solution, (10), which is sufficient for executing the algorithm.

2.3 Finding an Improved Basis

Given the results of section 2.2, we now address the question of which p vectors from N should become basic. Let the cost vector, $\bar{c}$, be partitioned as follows: $\bar{c} = [\bar{c}_{N1}, \bar{c}_{B2}, \bar{c}_{B1}, \bar{c}_{N2}]$. Then the new objective function may be written as follows:

(11)
$$\hat{z} = \bar{c}_{N1}'\hat{\bar{x}}_{B1} + \bar{c}_{B2}'\hat{\bar{x}}_{B2},$$

where $\bar{c}_{N1}' = [c_1, \dots, c_p]$ and $\bar{c}'_{B2} = [c_{p+1}, \dots, c_m]$. From (10) we have

$$\hat{\bar{x}}_{B1} = Y_{11}^{-1}\bar{x}_{B1} = [\hat{x}_{B1}, \hat{x}_{B2}, \dots, \hat{x}_{Bp}]$$
$$\hat{\bar{x}}_{B2} = \bar{x}_{B2} - Y_{21}Y_{11}^{-1}\bar{x}_{B1} = \bar{x}_{B2} - Y_{21}\hat{\bar{x}}_{B1}$$

and

(12)
$$\bar{x}_{B2} = [x_{B,p+1}, x_{B,p+2}, \dots, x_{Bm}]$$
$$Y_{21}\hat{\bar{x}}_{B1} = \sum_{j=1}^{p} y_{ij}\hat{x}_{Bj} \quad i = p+1, p+2, \dots, m.$$

Combining (11), (12) and the definitions of $\bar{c}_{N1}$, $\bar{c}_{B2}$, we have:

(13)
$$\hat{z} = \sum_{j=1}^{p} c_j\hat{x}_{Bj} + \sum_{i=p+1}^{m} c_{Bi}\left[x_{Bi} - \sum_{j=1}^{p} y_{ij}\hat{x}_{Bj}\right].$$

In (13) the terms:

(14)
$$x_{Bi} - \sum_{j=1}^{p} y_{ij}\hat{x}_{Bi}$$

are missing for $i = 1, 2, \dots, p$. However, for $i = 1, 2, \dots, p$ it is seen from (10) that

$$\bar{x}_{B1} = Y_{11}\hat{\hat{x}}_{B1}$$

or in component form:

$$x_{Bi} = \sum_{j=1}^{p} y_{i\,j}\hat{x}_{Bj} \qquad i = 1, 2, \dots, p.$$

Therefore the terms given in (14) are zero. Hence we may rewrite (13) as:

$$\hat{z} = \sum_{j=1}^{p} c_j\hat{x}_{Bj} + \sum_{i=1}^{m} c_{Bi}\left[x_{Bi} - \sum_{j=1}^{p} y_{i\,j}\hat{x}_{Bj}\right]$$

(15)
$$= \sum_{i=1}^{m} c_{Bi}x_{Bi} + \sum_{j=1}^{p}\left[c_j - \sum_{i=1}^{m} c_{Bi}y_{i\,j}\right]\hat{x}_{Bj}.$$

Note that:

(16)
$$\sum_{i=1}^{m} c_{Bi}x_{Bi} = z,$$

Let

(17)
$$\sum_{i=1}^{m} c_{Bi}y_{i\,j} = z_j \qquad j = 1, 2, \dots, p.$$

From (15), (16) and (17) we have:

(18)
$$\hat{z} = z + \sum_{j=1}^{p} (c_j - z_j)\hat{x}_{Bj}.$$

We see clearly from (18) that, providing the $\hat{x}_{Bj} \geqslant 0$, if we choose vectors to enter the basis for which all $z_j - c_j < 0$, $j = 1, 2, \dots, p$ then $\hat{z} \geqslant z$. Therefore the criterion for vectors to enter the basis is the same as it is for the single vector simplex method.

2.4 Feasibility Considerations

Let us now examine the issue of how to choose the vectors to leave the basis so that the new solution $\hat{\hat{x}}_B$ is feasible. The relations that must be satisfied are:

$$\hat{\hat{x}}_{B1} = Y_{11}^{-1}\,\bar{x}_{B1} \geqslant 0$$

(19)
$$\hat{\hat{x}}_{B2} = \bar{x}_{B2} - Y_{21}Y_{11}^{-1}\,\bar{x}_{B1} \geqslant 0.$$

We first examine the simplest case, namely the one for which $p = 2$.

Paranjape [5] studied this case and arrived at an incorrect set of criteria for the vectors to be removed from the basis. Paranjape maintains, in our notation, that the following criteria guarantee that the new basis and basic solution will be feasible:

1. The elements of Y_{11} must be non-negative and $\det(Y_{11}) > 0$.

2. x_{B1}, x_{B2} correspond to the removal of $\bar{b}_1$, $\bar{b}_2$, from the basis if:

$$\frac{x_{B1}}{y_{11}} = \min\left\{\frac{x_{Bi}}{y_{i1}} : y_{i1} > 0\right\}$$

$$\frac{x_{B2}}{y_{22}} = \min\left\{\frac{x_{Bi}}{y_{i2}} : y_{i2} > 0\right\}.$$

He chooses two vectors, say $\bar{a}_1$, $\bar{a}_2$ to enter such that $z_1 - c_1 < 0$, $z_2 - c_2 < 0$ are the most negative and then uses the above criteria to select the vectors to be removed. A simple counter example to Paranjape's method is as follows.

Suppose at some iteration of Paranjape's 2-variable method the basis is for convenience, an identity matrix, i.e.,

$$\bar{b}_1 = \begin{bmatrix}1\\0\\0\end{bmatrix}, \ \bar{b}_2 = \begin{bmatrix}0\\1\\0\end{bmatrix}, \ \bar{b}_3 = \begin{bmatrix}0\\0\\1\end{bmatrix}$$

$$\bar{a}_1 = \begin{bmatrix}3\\1\\2\end{bmatrix} = \begin{bmatrix}y_{11}\\y_{21}\\y_{31}\end{bmatrix}; \ \bar{a}_2 = \begin{bmatrix}1\\4\\2\end{bmatrix} = \begin{bmatrix}y_{12}\\y_{22}\\y_{32}\end{bmatrix}$$

$$\bar{b} = \begin{bmatrix}2\\3\\2\end{bmatrix} = \begin{bmatrix}x_{B1}\\x_{B2}\\x_{B3}\end{bmatrix}.$$

Paranjape's method would calculate:

$$\frac{x_{B1}}{y_{11}} = \min\left\{\frac{2}{3}, \frac{3}{1}, \frac{2}{2}\right\} = \frac{2}{3},$$

and

$$\frac{x_{B2}}{y_{12}} = \min\left\{\frac{2}{1}, \frac{3}{4}, \frac{2}{2}\right\} = \frac{3}{4}$$

since $Y_{11} = \begin{bmatrix}y_{11} & y_{12}\\ y_{21} & y_{22}\end{bmatrix} = \begin{bmatrix}3 & 1\\ 1 & 4\end{bmatrix}$, $\det(Y_{11}) = 11$. Hence, according to Paranjape, since $\det(Y_{11}) > 0$, the above criteria for removing $\bar{b}_1$, $\bar{b}_2$ from the basis and replacing them with $\bar{a}_1$, $\bar{a}_2$ should be correct. Therefore, the new basis is $\hat{b}_1 = \bar{a}_1$, $\hat{b}_2 = \bar{a}_2$, $\hat{b}_3 = \bar{b}_3$. The new basic solution is:

$$\hat{x}_{B1} = Y_{11}^{-1}\bar{x}_{B1} = \frac{1}{11}\begin{bmatrix}4 & -1\\ -1 & 3\end{bmatrix}\begin{bmatrix}2\\3\end{bmatrix} = \begin{bmatrix}\frac{5}{11}\\[2pt] \frac{7}{11}\end{bmatrix}.$$

$\hat{x}_{B2} = \bar{x}_{B2} - Y_{21}\hat{x}_{B1}$ and $\hat{x}_{B2}$ is the scalar $\hat{x}_{B3}$ for this problem. Therefore

$$\hat{x}_{B3} = 2 - (2,2)\begin{bmatrix}\frac{5}{11}\\[2pt] \frac{7}{11}\end{bmatrix} = \frac{-2}{11} < 0.$$

Hence Paranjape's method has led to an infeasible solution. The new solution is a *basic* solution since:

$$\frac{5}{11}\begin{bmatrix}3\\1\\2\end{bmatrix}+\frac{7}{11}\begin{bmatrix}1\\4\\2\end{bmatrix}-\frac{2}{11}\begin{bmatrix}0\\0\\1\end{bmatrix}=\begin{bmatrix}2\\3\\2\end{bmatrix}$$

but it is *not* feasible.

Blocher [1] gives a correct analysis of the two vector basis entry algorithm.

Let us suppose that a_p and a_q are to enter the basis and we wish to determine which vectors b_s, b_t should leave the basis. With this solution, the conditions for feasibility are:

$$\hat{x}_{Bs} = \theta_p = \frac{x_{Bs}y_{tq} - x_{Bt}y_{sq}}{y_{sp}y_{tq} - y_{sq}y_{tp}} \geqslant 0 \tag{20}$$

$$\hat{x}_{Bt} = \theta_q = \frac{x_{Bt}y_{sp} - x_{Bs}y_{tp}}{y_{sp}y_{tq} - y_{sq}y_{tp}} \geqslant 0 \tag{21}$$

$$\hat{x}_{Bi} = x_{Bi} - \theta_p y_{ip} - \theta_q y_{iq} \geqslant 0, \quad i \neq s,t \tag{22}$$

$$y_{sp}y_{tq} - y_{sq}y_{tp} \neq 0, \tag{23}$$

where (23) guarantees that $\text{Det}(Y_{11}) \neq 0$. Analysis for more than a two vector exchange is considerably more complicated and has not been attempted here.

2.5 The Algorithm

Given the results of the previous sections we now propose a multiple basis entry algorithm. We will restrict attention to a two vector exchange and use (23) to insure that Theorem 1 is satisfied. We take the view that checking (20)-(22) to determine the leaving variables is too heavy a computational burden and we propose to move to the new basic solution even if feasibility is violated. The use of a composite objective function (see [6]) will be used to drive the solution back to the feasible region. Furthermore, we will use the two vector replacement for only a fixed number of iterations and then use only single replacements thereafter, hence finiteness and convergence is guaranteed by the simplex method.

ALGORITHM-1: MULTIPLE BASIS EXCHANGE ALGORITHM

0. Initialization

Let $[\bar{x}_B \,\vdots\, \bar{x}_N]$ be a basic solution. Save the objective function, $\bar{c}^* \leftarrow \bar{c}$, select β for the composite objective function, and set γ to the desired value for the maximum number of double basis replacement iterations. Set $i \leftarrow 0$.

1. Composite Objective

$$\text{Set } d_j \leftarrow \begin{cases} 0, & \text{if } x_j \geqslant 0 \\ 1, & \text{if } x_j < 0. \end{cases}$$

Set $\bar{c} \leftarrow \bar{c}^* + \beta\bar{d}$.

2. Pricing

Let $\psi = \{x_j \text{ nonbasic: } z_j - c_j < 0\}$. If $\psi = \phi$, terminate with $[\bar{x}_B \vdots \bar{x}_N]$ optimal. If $i < \gamma$ and $|\psi| \geqslant 2$, then select x_p and x_q from ψ to enter the basis and set $\alpha \leftarrow 2$; otherwise, select x_p from ψ to enter the basis and set $\alpha \leftarrow 1$.

3. Select Leaving Variable

a. If $\alpha = 2$, select two vectors, say $\bar{b}_s$, $\bar{b}_t$, to leave the basis; otherwise, go to *c.*

b. If $y_{sp}y_{tq} - y_{sq}y_{tp} \neq 0$, go to 4; otherwise, go to *c.*

c. Let only x_p enter the basis, perform a simplex pivot, set $i \leftarrow i+1$ and return to 1.

4. Update

$$\text{Set } B^{-1} \leftarrow \left[\begin{array}{c|c} Y_{11}^{-1} & \\ \hline -Y_{21}Y_{11}^{-1} & I \end{array}\right] B^{-1},$$

$\bar{x}_{B1} \leftarrow Y_{11}^{-1}\bar{x}_{B1}$, and

$\bar{x}_{B2} \leftarrow \bar{x}_{B2} - Y_{21}Y_{11}^{-1}\bar{x}_{B1}$. Set $i \leftarrow i + 1$ and go to 1.

An example problem illustrating ALGORITHM-1 is presented in Appendix A.

3. A FEASIBLE DIRECTIONS ALGORITHM

For the algorithm of Section 2 we allowed one (or two) nonbasic variables to increase in value at each iteration and required that the same number of basic variables assume the new value of zero. In this section we relax the requirement that the nonbasic variables must assume the value of zero, but we require that feasibility be maintained.

In developing this algorithm we have drawn freely from the theory of nonlinear programming and the development which follows is simply our implementation of a feasible direction procedure. Other related ideas may be found in [2].

For generality we rewrite the linear program with upper bound constraints as follows:

$$(24) \qquad \left.\begin{array}{c} \max \bar{c}'\bar{x} \\ \text{subject to } A\bar{x} = \bar{b} \\ \bar{0} \leqslant \bar{x} \leqslant \bar{u}. \end{array}\right|$$

Let B be any feasible basis for (24). Partitioning all other quantities in (24) into the basic and nonbasic components, we may rewrite (24) as follows:

$$(25) \qquad \max z = \bar{c}'_B\bar{x}_B + \bar{c}'_N\bar{x}_N$$

$$(26) \qquad \text{subject to } B\bar{x}_B + N\bar{x}_N = \bar{b}$$

$$(27) \qquad \bar{0} \leqslant \bar{x}_B \leqslant \bar{u}_B$$

(28) $$0 \leqslant \bar{x}_N \leqslant \bar{u}_N.$$

Solving (26) for $\bar{x}_B$ and substituting into (25) and (27) we obtain

$$\max z = (\bar{c}'_N - \bar{c}'_B B^{-1} N)\, \bar{x}_N + \bar{c}'_B B^{-1} \bar{b}$$

$$\text{subject to } \bar{0} \leqslant B^{-1}\bar{b} - B^{-1} N \bar{x}_N \leqslant \bar{u}_B$$

$$\bar{0} \leqslant \bar{x}_N \leqslant \bar{u}_N.$$

We define any direction in nonbasic space as a *feasible direction of movement.* This is because

$$\bar{x}_B = B^{-1}\bar{b} - B^{-1} N \bar{x}_N$$

may be used to determine $\bar{x}_B$ given any $\bar{x}_N$. An *improving feasible direction* at a nonbasic point $\bar{x}_N^0$ is any direction $\bar{d}$ having $\nabla z(\bar{x}_N^0) \cdot \bar{d} > 0$, where $\nabla z(\bar{x}_N^0)$ is the gradient of z evaluated at $\bar{x}_N^0$. Since we are dealing with a linear function, $\nabla z(\bar{x}_N^0) = \bar{c}'_N - \bar{c}'_B B^{-1} N$. The maximum movement in some direction $\bar{d}$ (i.e., $\bar{x}_N + \alpha\bar{d}$) is restricted by (27) and (28). Then α is restricted by

(29) $$B^{-1}\bar{b} - B^{-1} N \bar{x}_N - \bar{u}_B \leqslant B^{-1} N\, \alpha\, \bar{d} \leqslant B^{-1}\bar{b} - B^{-1} N \bar{x}_N$$

and

(30) $$-\bar{x}_N \leqslant \alpha\, \bar{d} \leqslant \bar{u}_N - \bar{x}_N$$

We now present a general feasible directions algorithm for the linear program (24).

ALGORITHM-2: FEASIBLE DIRECTIONS ALGORITHM

0. Initializaton

Find a feasible basis B with corresponding solution $[\bar{x}_B, \bar{x}_N]$. Let $\epsilon > 0$ be some termination tolerance.

1. Pricing

Select a direction vector $\bar{d}$ such that

$$d_j \leftarrow \left\{\begin{array}{ll} 1, & \text{if } x_{Nj} < u_N \text{ and } c_j - z_j > \epsilon. \\ -1, & \text{if } x_{Nj} > 0 \text{ and } c_j - z_j < \epsilon. \\ 0, & \text{otherwise.} \end{array}\right\}$$

If $\bar{d} = \bar{0}$, terminate.

2. Maximum Movement

Find the largest α, say α^*, such that (29) and (30) are satisfied.

3. Update Variables

Let

$$\bar{x}_N \leftarrow \bar{x}_N + \alpha^* \bar{d} \text{ and}$$

$$\bar{x}_B \leftarrow \bar{x}_B - B^{-1} N\, \alpha^* \bar{d}$$

4. Pivot Required

If α^* is determined by (30) return to step 1; otherwise, at least one basic variable has been driven to either zero or its upper bound. Replace this basic variable, say x_{Br}, by some nonbasic variable, say x_{Nj}, having $y_{rj} \neq 0$ where $y_r = B^{-1}\bar{a}_j$.

Note that if in step 1 we require that $\bar{d}$ have exactly one nonzero element, then the above algorithm becomes precisely the primal simplex method. To our knowledge, the computational efficiency of this type of algorithm has never been experimentally studied. An example problem has been solved using ALGORITHM-2 in Appendix B.

APPENDIX A
Example Using ALGORITHM-1

Consider the following example

$$\begin{array}{ll} \max & x_1 + x_2 \\ \text{s.t.} & 2x_1 + 3x_2 \leqslant 6 \\ & x_1 - 2x_2 \leqslant 1 \\ & x_1 \leqslant 2 \\ & x_1,\ x_2 \geqslant 0 \end{array} \Rightarrow \begin{array}{ll} \max & x_1 + x_2 \\ \text{s.t.} & 2x_1 + 3x_2 + x_3 = 6 \\ & x_1 - 2x_2 + x_4 = 1 \\ & x_1 + x_5 = 2 \\ & x_1,\ x_2,\ x_3,\ x_4,\ x_5 \geqslant 0 \end{array}$$

0. (Initialization)

   Let $\left[x_1^N\ x_2^N\ x_3^B\ x_4^B\ x_5^B\right] = [0\ 0\ 6\ 1\ 5]$.

   Set $\bar{c}^* = [1\ 1\ 0\ 0\ 0]$, $\beta = 2$, $\gamma = 1$, $i = 0$.

1. (Composite Objective)

   $\bar{d} = [0\ 0\ 0\ 0\ 0]$ and $\bar{c} = [1\ 1\ 0\ 0\ 0]$

2. (Pricing)

   x_1: $z_1 - c_1 = -1 \Rightarrow x_1 \in \psi$

   x_2: $z_2 - c_2 = -1 \Rightarrow x_2 \in \psi$

   Let $x_p = x_1$, $x_q = x_2$, and $\alpha = 2$.

3. (Select Leaving Variable)

   Let $[\bar{b}_s,\ \bar{b}_t] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$ Then

   $Y_{11} = \begin{array}{cc|c} s & t & \\ 2 & 3 & p \\ 1 & -2 & q \end{array}$ and $y_{sp}y_{tq} - y_{sq}y_{tp}$

   $= (2)\ (-2) - (1)\ (3) = -7$

4. (Update)

$$Y_{11}^{-1} = \begin{bmatrix} 2/7 & 3/7 \\ 1/7 & -2/7 \end{bmatrix} \text{ and } Y_{21} = [1\ 0].$$

$$B^{-1} = \left[\begin{array}{cc|c} 2/7 & 3/7 & \\ 1/7 & -2/7 & \\ \hline -2/7 & -3/7 & 1 \end{array}\right] \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} = \begin{bmatrix} 2/7 & 3/7 & 0 \\ 1/7 & -2/7 & 0 \\ -2/7 & -3/7 & 1 \end{bmatrix},$$

$$\bar{x}_{B1} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2/7 & 3/7 \\ 1/7 & -2/7 \end{bmatrix} \begin{bmatrix} 6 \\ 1 \end{bmatrix} = \begin{bmatrix} 15/7 \\ 4/7 \end{bmatrix}$$

$$\bar{x}_{B2} = x_5 = 2 - [2/7\ \ 3/7] \begin{bmatrix} 6 \\ 1 \end{bmatrix} = -1/7.$$

$i = 2.$

1. (Composite Objective) $[x_1^B\ x_2^B\ x_3^N\ x_4^N\ x_5^B] = [15/7\ 4/7\ 0\ 0\ -1/7]$

   $\bar{d} = [0\ 0\ 0\ 0\ 1]$ and $\bar{c} = [1\ 1\ 0\ 0\ 2]$.

2. (Pricing)

   x_3: $z_3 - c_3 = -1/7 \Rightarrow x_3 \in \psi$

   x_1: $z_4 - c_4 = -5/7 \Rightarrow x_4 \in \psi$

   Let $x_p = x_4$ and $\alpha = 1$.

3. (Select Leaving Variable)

Leaving variable is x_5

$$B^{-1} = \begin{bmatrix} 0 & 0 & 1 \\ 1/3 & 0 & -2/3 \\ 2/3 & 1 & -7/3 \end{bmatrix}$$

$$[x_1^B\ x_2^B\ x_3^N\ x_4^B\ x_5^N] = [2\ \ 2/3\ \ 0\ \ 1/3\ \ 0]$$

1. (Composite Objective)

   $\bar{d} = \bar{0}$, $\bar{c} = [1\ 1\ 0\ 0\ 0]$

2. (Pricing)

   x_3: $z_3 - c_3 = 1/3$

   x_5: $z_5 - c_5 = 1/3$

   $\Rightarrow$ Optimality!

The sequence of pivots is illustrated in Figure 1.

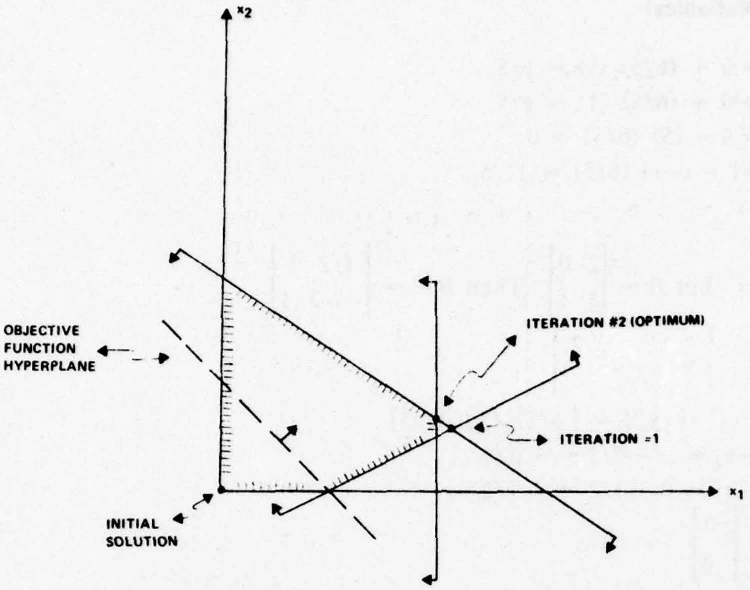


FIGURE 1. Example problem solution by ALGORITHM-1.

APPENDIX B Example Using ALGORITHM-2

Consider the following example

$$\begin{array}{llcll} & \max x_1 + x_2 & & & \max x_1 + x_2 \\ s.t. & 2x_1 + 3x_2 \leqslant 6 & & s.t. & 2x_1 + 3x_2 + x_3 = 6 \\ & x_1 - 2x_2 \leqslant 1 & \rightarrow & & x_1 - 2x_2 + x_4 = 1 \\ & 0 \leqslant x_1 \leqslant 2 & & & 0 \leqslant x_1 \leqslant 2, \\ & 0 \leqslant x_2 \leqslant \infty & & & 0 \leqslant x_2,\ x_3,\ x_4 \leqslant \infty. \end{array}$$

0. (Initialization)

Let $[x_1^N\ x_2^N\ x_3^B\ x_4^B\ ] = [0\ \ 0\ \ 6\ \ 1]$, and $\epsilon = 0$.

1. (Pricing)

$c_1 - z_1 = 1,\ \ c_2 - z_2 = 1.$

$$\bar{d} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

2. (Maximum Movement)

$$\left.\begin{array}{c} -\begin{bmatrix} \infty \\ \infty \end{bmatrix} \leqslant \begin{bmatrix} 5 \\ -1 \end{bmatrix} \alpha \leqslant \begin{bmatrix} 6 \\ 1 \end{bmatrix} \\ \text{and} \\ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \leqslant \begin{bmatrix} 1 \\ 1 \end{bmatrix} \alpha \leqslant \begin{bmatrix} \infty \\ \infty \end{bmatrix} \end{array}\right\} \rightarrow \alpha \leqslant 6/5$$

3. (Update Variables)

$$\begin{aligned}
x_1 &= 0 + (6/5)\ (1) = 6/5\\
x_2 &= 0 + (6/5)\ (1) = 6/5\\
x_3 &= 6 - (5)\ (6/5) = 0\\
x_4 &= 1 - (-1)\ (6/5) = 11/5
\end{aligned}$$

4. (Pivot)

$$\text{Let } B = \begin{bmatrix} 2 & 0\\ 1 & 1 \end{bmatrix}. \text{ Then } B^{-1} = \begin{bmatrix} 1/2 & 0\\ -1/2 & 1 \end{bmatrix}.$$

1. (Pricing)

$$\begin{aligned}
&[x_1^B\ x_2^N\ x_3^N\ x_4^B] = [\ 6/5\ 6/5\ 0\ 11/5]\\
&c_2 - z_2 = 1 - 3/2 = -\ 1/2\\
&c_3 - z_3 = 0 - 1/2 = -\ 1/2\\
&\bar{d} = \begin{bmatrix} -1\\ 0 \end{bmatrix}
\end{aligned}$$

2. (Maximum Movement)

$$\left.\begin{aligned}
&\begin{bmatrix} 3\\ -2 \end{bmatrix} - \begin{bmatrix} 9/5\\ -21/5 \end{bmatrix} - \begin{bmatrix} 2\\ \infty \end{bmatrix} \leqslant \begin{bmatrix} -3/2\\ 7/2 \end{bmatrix} \alpha \leqslant \begin{bmatrix} 3\\ 2 \end{bmatrix} - \begin{bmatrix} 9/5\\ -21/5 \end{bmatrix}\\
&\qquad\qquad\qquad\qquad \text{and}\\
&\qquad\qquad \begin{bmatrix} -6/5\\ 0 \end{bmatrix} \leqslant \begin{bmatrix} -1\\ 0 \end{bmatrix} \alpha \leqslant \begin{bmatrix} \infty\\ \infty \end{bmatrix}
\end{aligned}\right\} \rightarrow \alpha \leqslant 8/15$$

3. (Update Variables)

$$\begin{aligned}
x_2 &= 6/5 + 8/15(-1) = 2/3\\
x_3 &= 0 + (8/15)\ (0) = 0\\
x_1 &= 6/5 - (8/15)\ (-3/2) = 2\\
x_4 &= 11/5 - (8/15)\ (7/2) = 1/3
\end{aligned}$$

4. (Pivot)

$$\text{Let } B = \begin{bmatrix} 3 & 0\\ -2 & 1 \end{bmatrix} \quad 3^{-1} = \begin{bmatrix} 1/3 & 0\\ 2/3 & 1 \end{bmatrix}$$

1. (Pricing)

$$\begin{aligned}
c_1 - z_1 &= 1 - 2/3 = 1/3\\
c_3 - z_3 &= 0 - 1/3 = -1/3
\end{aligned}$$

Terminate!

The solution sequence is illustrated in Figure 2.

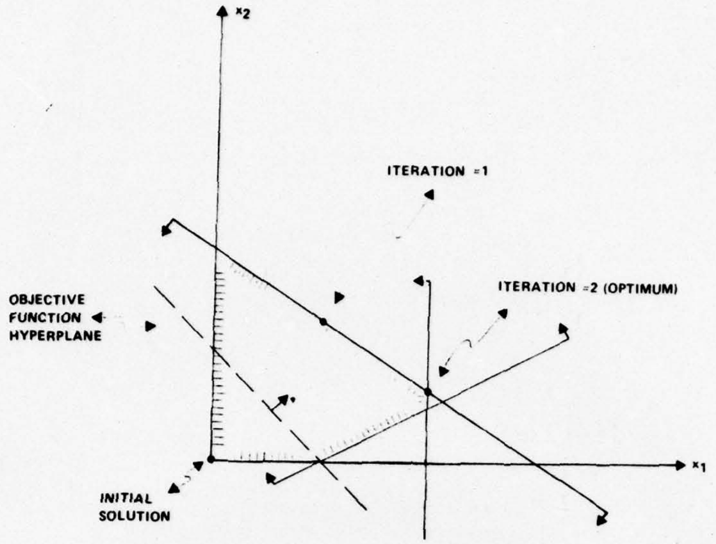


FIGURE 2. Example problem solution by ALGORITHM-2.

REFERENCES

[1] Blocher, R. H., "The Revised Duplex Algorithm for Linear Programming," M. S. Thesis, Washington University, (1966).

[2] Brown, G. W., and T. C. Koopmans, "Computational Suggestions for Maximizing a Linear Function Subject to Linear Inequalities," pp. 377-380, *Activity Analysis of Production and Allocation,* Edited by T. K. Koopmans (John Wiley and Sons, New York, 1951).

[3] Fulkerson, D. R., "An Out-of-Kilter Method for Minimal-Cost Flow Problems," Journal of the Society of Industrial and Applied Mathematics, 9(1), 18-27 (1961).

[4] Kuhn, H. W., "The Hungarian Method for the Assignment Problem," Naval Research Logistics Quarterly, 2, 83-97 (1955).

[5] Paranjape, S. R., "The Simplex Method: Two Basic Variables Replacement," Management Science, 12, 135-141, (1965).

[6] Wolfe, P., "The Composite Simplex Algorithm," SIAM Review, 7(1), 42-54, (1965).

DESIGN OF A PROCESS CONTROL SCHEME FOR DEFECTS PER 100 UNITS BASED ON AOQL*

Richard S. Leavenworth and Richard L. Scheaffer

University of Florida
Gainesville, Florida

ABSTRACT

A process control scheme is developed in which decisions as to the frequency of sampling are made based upon the choice of an Average Outgoing Quality Limit. The scheme utilizes plotted points on a U-control chart for defects and the theory of runs to determine when to switch among Reduced, Normal, Tightened, and 100 percent inspection. The scheme is formulated as a semi-Markov process to derive steady state equations for the probabilities of being in Reduced, Normal, Tightened, or 100 percent inspection and for Average Outgoing Quality and Average Fraction Inspected. The resulting system and the computer programs used to derive it are discussed.

INTRODUCTION

In current process control an important question that still appears not to have found an adequate answer is "How much inspection is enough inspection?" Standard texts in process quality control tend to sidestep this issue. Frequently examples are given in which control charting inspection data is acquired in subgroups under 100 percent inspection. This is more likely to be the case during new process start-up. Other examples, usually cases of continued use of control charts, indicate some form of sampling. Seldom is there mention made of the decision process that determines at what point a change is made from one procedure to the other.

In more recent years research on process control methodologies has turned to cost-based economic models (frequently based on some Bayesian notions about physical operation of the process) to help provide an analytical framework within which to answer this most important question of resource justification and allocation. To mention only two such developments in the area of process control, Duncan [2], proposed an economic design for the $\bar{X}$ chart in 1956, and Montgomery, Heikes, and Mance [7] proposed a p-chart model in 1975.

For one reason or another, there is little indication in the literature that these models have found application in industry. Nor do they appear to have been picked up by any of the standard texts in process quality control.

*This research was supported by the U.S. Department of the Navy, Office of Naval Research under Contracts Number N00014-68-A-0173-0021 and N00014-75-C-0783.

The interest of the authors in this question was stimulated when they were contacted by quality control representatives from a large aircraft maintenance facility. Without the manufacturing activity, the usual continuous production lines and/or job shop operation producing batches of essentially identical items did not exist. Thus, a reasonable approach seemed to be to attempt to control maintenance operations on the basis of mistakes made per 100 man-hours worked rather than on the basis of percent defective items.

The general approach to the problem chosen by the authors was to adapt the concept of Acceptance Control Charting, developed by Freund [3], and the AOQL concept from acceptance sampling, developed by Dodge [1], to the basic u-chart model. Freund based his $\bar{X}$ chart model on the selection of two points on the (OC) Operating Characteristic Curve, the Acceptable Process Level (APL), with Type I error probability α, and a Rejectable Process Level (RPL), with Type II error probability β. Solution of the model yields a sample subgroup size and a one-sided control limit. The Freund formulation, however, does not yield an analytical solution to the question, "How much inspection is enough inspection?"

First developed for lot-by-lot acceptance sampling by attributes and later used as the foundation of continuous sampling, Dodge's AOQL concept of rectifying inspection is based on alternating back and forth from 100% inspection to sampling inspection. Thus, over the long haul, the average quality will be a mix of product, part of which is some proportion defective and part of which is (presumably) perfect. It is precisely this aspect of rectifying inspection that the authors chose as the analytical base upon which to answer the "How much" question.

APPLICATION

Interest in attacking the problem of in-process inspection was originally stimulated by a perceived need on the part of a large aircraft maintenance facility. Therefore, many of the system design assumptions, development, and basic formulations that follow are based on the specific needs and operating methods of that organization.

The organization was familiar with acceptance sampling procedures and in particular those of MIL-STD-105D [6] and Handbook H-106 on multi-level continuous sampling. Sampling inspection was used in shops processing non-critical items. Sample sizes, applicable on a monthly basis, were selected based on a modification of MIL-STD-105D. However, defective items found in excess of the MIL-STD-105D acceptance numbers led to no positive action because there was no formed lot upon which action could be taken.

The basic inspection unit used in this process control scheme is the manhour of productive work. Sampling is based on inspecting units of output and recording the time required to complete the work and the number of defects found. The time span during which a sample of some proportion of the total manhours worked will be drawn, if the shop is not on 100% inspection, is called a Production Interval. In the theoretical developments that follow, it is assumed that the inspection process never passes over a defect. That is, when 100% inspection is in force, all defects are intercepted and repaired or replaced, and that defects found in samples also are intercepted.

In describing the various aspects of this study, we have adopted the terminology used by Hill [5] and expanded upon by Stephens and Larson [9]. The term sampling *scheme* refers to "an over-all strategy specifying the way in which sampling plans are to be used." Under this definition, MIL-STD-105D is an AQL (Acceptable Quality Level) scheme; discussed herein is a combination AQL-AOQL (Average Outgoing Quality Limit) scheme. The term sampling *plan* is used to describe "the specification of the rules to be followed in sentencing any particular batch of articles," or, in this case, sentencing the process itself. The term sampling *system* refers to the group of sampling plans and the switching rules used to determine when to switch from one plan to another and to and from 100% inspection.

Ground Rules for Sampling Scheme

The basic ground rules under which the procedure was developed, largely established because of the company organization's familiarity with MIL-STD-105D, were as follows:

(1) The concept of AQL, established and long accepted in government standards and specifications, would be maintained. The AQL is one basis for coding the various sampling plans in the inspection scheme.

(2) Production Interval ranges, in terms of man-hours worked during a fixed period of time, and standard AQL values would increase in a geometric pattern in accordance with the procedures employed to set preferred AQL values in MIL-STD-105D, roughly multiples of $10^{0.2}$.

(3) Provisions for normal, tightened, and reduced inspection would be included as well as provisions for 100% inspection. It is the provision for 100% inspection that allows calculation of an AOQL.

(4) Emphasis was to be on process control; therefore, decision rules focus on control charting techniques.

(5) Procedures would need to be simple and straight-forward and require as little calculation as possible.

(6) AOQ (Average Outgoing Quality) functions were to be provided. When rectifying inspection is employed, the AOQL provides a measure of the level of protection to the customer. The scheme was to be coded against the AOQL as well as the AQL.

(7) Shop history of defects, AQL levels, and man-hour levels determine the amount of inspection required. AFI (Average Fraction Inspected) curves were to be provided in order to assist management in projecting required inspection manpower levels.

Measures of Effectiveness

The measures of effectiveness considered for the scheme included the Average Outgoing Quality Limit and the Average Fraction Inspected (AFI). The scheme is coded against an AQL and an AOQL in which three different levels of inspection are provided dependent upon management's requirement. They are designated Inspection Levels I, II, and III, in order of increasing stringency, respectively. In each case:

$$\text{AOQL} = \text{AQL}\ (10^{0.2y})$$

where y = 2, for Inspection Level I
1, for Inspection Level II
0, for Inspection Level III

(AOQL = AQL for all sampling systems designated under Inspection Level III).

This relationship was chosen for reasons of practicality and convenience. The theory would work with any prescribed relationship.

AFI, defined as the proportion of the productive man-hours subjected to inspection, was chosen as the objective function to be minimized. This measure is converted into an average number of items checked by the Q. C. specialist based on the number of man-hours worked in a shop during a designated Production Interval and the number of items processed during the period. During any given Production Interval, actual hours included in a sample may vary from the planned amount.

Since AFI ranges from the proportion of hours inspected under reduced inspection to 1, when consistently on 100 percent inspection, it is necessary to identify some specific value of production quality at which the AFI is to be minimized. The point chosen was the AQL, that rate of defect production considered to be the maximum allowable as a process average.

SYSTEM DESCRIPTION

SOME DEFINITIONS

Definitions of terms and symbols required for discussion of the rectifying inspection system follow.

AFI = Average Fraction Inspected. The proportion of total production of a shop which will be inspected, on the average, when the rate of defects is at a stipulated level.

AOQ = Average Outgoing Quality. The average quality of outgoing production in defects per 100 man-hours when the rules and procedures of the sampling system are applied.

AOQL = Average Outgoing Quality Limit. The maximum value that average outgoing quality, in defects per 100 man-hours, will reach when the rules and procedures of the sampling system are applied.

AQL = Acceptable Quality Level. The maximum rate of defect production, in defects per 100 man-hours, which, for purposes of sampling inspection, can be considered acceptable as a process average.

CL = Control Limit (Upper) for a Shop Control Chart for defects per 100 man-hours.

CL_r = Control Limit under reduced sampling inspection.

CL_n = Control Limit under normal sampling inspection and under 100% inspection.

CL_t = Control Limit under tightened sampling inspection.

H = Production Interval. The number of shop production man-hours worked during a pre-determined period of time such as a day, a week, or a month. In normal practice the week is chosen as the Production Interval.

h_n = Number of man-hours of production to be included in the sample during one production interval on the normal sampling inspection. The fraction of productive time is given by $f_n = h_n/H$. During 100 percent inspection, inspection results are accumulated in subgroups of h_n hours.

h_r = Number of man-hours of production to be included in the sample during one production interval when on the reduced sampling inspection. The fraction of productive time is given by $f_r = h_r/H$.

h_t = Number of man-hours of production to be included in the sample during one production interval when on the tightened sampling inspection. The fraction of productive time is given by $f_t = h_t/H$.

Sampling System Design

The generalized form of the sampling system is shown in Figure 1. The parameters which were varied for analytical purposes include:

(1) The sampling rates f_n, f_r, and f_t.

(2) The run lengths required to produce a switch among reduced, normal, tightened, and 100 percent inspection. These are designated k_1, k_2, k_3, k_4, and k_5.

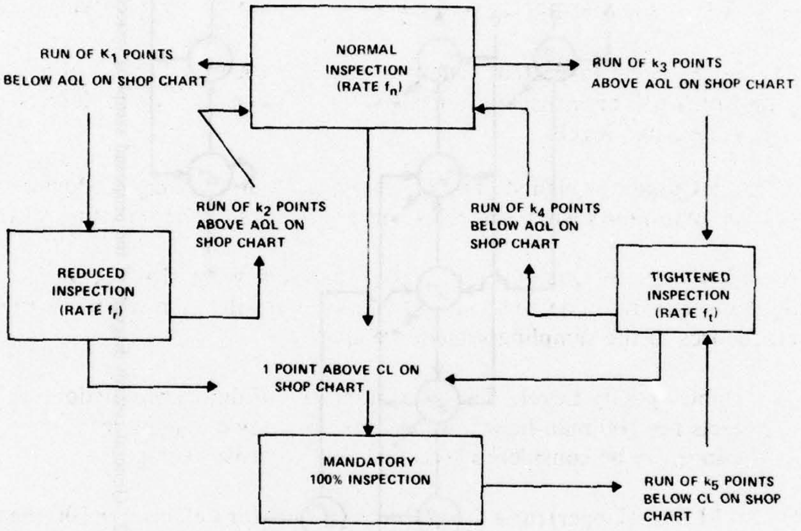

RUN OF K1 POINTS BELOW AQL ON SHOP CHART
NORMAL INSPECTION (RATE fn)
RUN OF k3 POINTS ABOVE AQL ON SHOP CHART
RUN OF k2 POINTS ABOVE AQL ON SHOP CHART
RUN OF k4 POINTS BELOW AQL ON SHOP CHART
REDUCED INSPECTION (RATE fr)
TIGHTENED INSPECTION (RATE ft)
1 POINT ABOVE CL ON SHOP CHART
MANDATORY 100% INSPECTION
RUN OF k5 POINTS BELOW CL ON SHOP CHART

FIGURE 1. Flow chart of generalized sampling system used for analysis

Thus the system provides for three different sampling intensities. Formal rules are provided for switching from one to another and to and from 100 percent inspection. When on 100 percent inspection, inspection data are accumulated into subgroups of h_n man-hours for control charting purposes and for the purpose of determining the time at which to return to sampling inspection. All decisions are based on the interpretation of inspection data points on a shop control chart.

A general system flow diagram is shown in Figure 2. This diagram was used to develop the probability transition matrix shown in Figure 3.

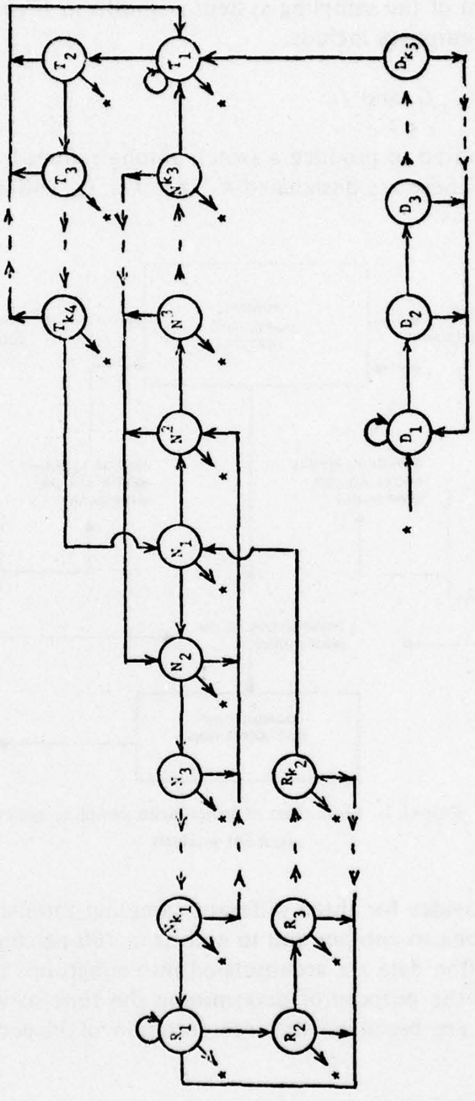


FIGURE 2. General system diagram of the proposed sampling procedure

Entering States

Leaving States		←Reduced→					←Normal→									←Tightened→					←Detailing→				
		R_{k_2}	R_{k_2-1}	...	R_2	R_1	N_{k_1}	N_{k_1-1}	...	N_2	N_1	N^2	...	N^{k_3-1}	N^{k_3}	T_1	T_2	...	T_{k_4-1}	T_{k_4}	D_1	D_2	...	D_{k_5-1}	D_{k_5}
←Reduced→	R_{k_2}	0	0	0	0	$P_\ell(r)$	0	0	0	0	$P_h(r)$	0	0	0	0	0	0	0	0	0	$1-P_\ell(r)$ $-P_h(r)$	0	0	0	0
	R_{k_2-1}	$P_h(r)$	0	.	.	$P_\ell(r)$	.	.	.	.	0	.	.	.	.	.	.	.	.	.	$1-P_\ell(r)$ $-P_h(r)$	.	.	.	.
	⋮	0	...	0	.	...	.	.	.	.	.	.	.	.	.	.	.	.	.	.	...	.	.	.	.
	R_2	.	0	...	0	$P_\ell(r)$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	$1-P_\ell(r)$ $-P_h(r)$	.	.	.	.
	R_1	.	.	0	$P_h(r)$	$P_\ell(r)$	.	.	.	.	.	0	.	.	.	.	.	.	.	.	$1-P_\ell(r)$ $-P_h(r)$	.	.	.	.
←Normal→	N_{k_1}	.	.	.	0	$P_\ell(n)$	0	.	.	.	.	$P_h(n)$	.	.	.	.	.	.	.	.	$1-P_\ell(n)$ $-P_h(n)$	.	.	.	.
	N_{k_1-1}	.	.	.	.	0	$P_\ell(n)$	0	.	.	.	$P_h(n)$	.	.	.	.	.	.	.	.	$1-P_\ell(n)$ $-P_h(n)$	.	.	.	.
	⋮	.	.	.	.	.	0	...	0	.	.	...	.	.	.	.	.	.	.	.	...	.	.	.	.
	N_2	.	.	.	.	.		0	...	0	.	$P_h(n)$	.	.	.	.	.	.	.	.	$1-P_\ell(n)$ $-P_h(n)$	.	.	.	.
	N_1	.	.	.	.	.	.	.	0	$P_\ell(n)$	.	$P_h(n)$	0	.	.	.	.	.	.	.	$1-P_\ell(n)$ $-P_h(n)$	.	.	.	.
	N^2	.	.	.	.	.	.	.	0	$P_\ell(n)$	.	0	...	0	.	.	.	.	.	.	$1-P_\ell(n)$ $-P_h(n)$	.	.	.	.
	⋮	.	.	.	.	.	.	.	.	...	.	.	0	...	0	.	.	.	.	.	...	.	.	.	.
	N^{k_3-1}	.	.	.	.	.	.	.	.	$P_\ell(n)$	.	.	.	0	$P_h(n)$	0	.	.	.	.	$1-P_\ell(n)$ $-P_h(n)$	.	.	.	.
	N^{k_3}	.	.	.	.	.	.	.	.	$P_\ell(n)$	.	.	.	.	0	$P_h(n)$	0	.	.	.	$1-P_\ell(n)$ $-P_h(n)$	.	.	.	.
←Tightened→	T_1	.	.	.	.	.	.	.	.	0	.	.	.	.	.	$P_h(t)$	$P_\ell(t)$	0	.	.	$1-P_\ell(t)$ $-P_h(t)$	.	.	.	.
	T_2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	$P_h(t)$	0	...	0	.	$1-P_\ell(t)$ $-P_h(t)$	.	.	.	.
	⋮	.	.	.	.	.	.	.	.	.	.	.	.	.	.	...	.	0	...	0	...	.	.	.	.
	T_{k_4-1}	.	.	.	.	.	.	.	.	.	0	.	.	.	.	$P_h(t)$	.	.	0	$P_\ell(t)$	$1-P_\ell(t)$ $-P_h(t)$	.	.	.	.
	T_{k_4}	.	.	.	.	.	.	.	.	.	$P_\ell(t)$	.	.	.	.	$P_h(t)$	.	.	.	0	$1-P_\ell(t)$ $-P_h(t)$	0	.	.	.
←Detailing→	D_1	.	.	.	.	.	.	.	.	.	0	.	.	.	.	0	.	.	.	.	$1-p(d)$	$p(d)$	0	.	.
	D_2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	$1-p(d)$	0	...	0	.
	⋮	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	...	.	0	...	0
	D_{k_5-1}	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	$1-p(d)$	.	.	0	$p(d)$
	D_{k_5}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$p(d)$	0	0	0	0	$1-p(d)$	0	0	0	0

FIGURE 3. Probability transition matrix for general system

State Transition Probabilities Defined

The following notation has been used to define the transition probabilities of Figure 3:

	Inspection States		
	Reduced	Normal	Tightened
Probability of a sample point below the AQL central line on the control chart.	$p_l(r)$	$p_l(n)$	$p_l(t)$
Probability of a point between the AQL and the CL.	$p_h(r)$	$p_h(n)$	$p_h(t)$
Probability of a point above CL.	$p_d(r)$	$p_d(n)$	$p_d(t)$

$p(d)$ = probability of a point below the CL (in control) on the control chart when on 100% inspection (detailing).

$= p_l(n) + p_h(n)$

$1 - p(d) =$ probability of a point above the CL (out of control) when on 100% inspection.

Formulation of Operating Characteristics of the System Model

The theoretical basis for developing formulas for the operating characteristics of the sampling system is given in Appendix A. The system model of Figures 1 and 2 was formulated as a semi-Markov process. Steady state probabilities of being in Reduced $P(R)$; Normal, $P(N)$: Tightened $P(T)$; and Detailing, $P(D)$; inspection were formulated in general from the probability transition matrix, Figure 3. These equations are given in Appendix B.

Operating Characteristic Curves were developed from the steady-state time-weighted probabilities, $P'(\cdot)$, of being in the various inspection states R, N, T, and D, given that the process is operating in control at specified levels of θ, the parametric rate of defects resulting per 100 man-hours worked. These are the likelihood functions, $L(\theta)$, or equivalently, the trace of $P'(\cdot)$ as the parameter θ is varied from 0 over its effective range.[1]

The time-weighted steady-state probabilities of being in the various states were then used in the equations for the measures of system effectiveness, AFI and AOQ.

From Appendix A, the development of equation A.5 yielded:

$$(1) \qquad P'(R) = \frac{HP(R)}{h_n P(D) + H[1 - P(D)]} = \frac{P(R)}{f_n P(D) + [1 - P(D)]}.$$

Corresponding equations for states N and T are equivalent to (1) with R replaced by N and T, respectively. For the detailing state, D,

[1] This is consistent with the binominal assumption used in producing the O.C. Curves for the various sampling plans contained in MIL-STD-105D [6] wherein it is assumed that the incoming fraction defective, p, is constant and results from a process operating in control at the level p. The O.C. Curves, then, are likelihood functions, $L(p)$, resulting from the trace of the probability of acceptance of a given lot as p is varied from 0 over its effective range ($\leqslant 1$). The same assumption was employed by Stephens and Larson [9] in their analysis of the operation of MIL-STD-105D as an acceptance sampling system.

(2)
$$P'(D) = \frac{f_n P(D)}{f_n P(D) + [1 - P(D)]}$$

The equation for Average Fraction Inspected also is derived in Appendix A, equation A.4. By eliminating *H* from this equation, we obtain:

$$\text{AFI} = [f_n P(D) + (1 - P(D))]^{-1} \, [f_r P(R) + f_t P(T) + f_n \{P(D) + P(N)\}]$$

which may also be expressed as:

(3)
$$\text{AFI} = f_r P'(R) + f_n P'(N) + f_t P'(T) + P'(D)$$

Since the time-weighted steady-state probabilities are required to produce the O.C Curves, this latter form of the equation is more useful.

The AOQ function is obtained from:

(4)
$$\text{AOQ} = \theta(1 - \text{AFI})$$

Calculation of Zone Probabilities

Use of the control statistic "defects per 100 man-hours" suggest a Poisson distribution of defects over time and the use of the U-chart for control charting purposes. The zones of interest on the U-chart, as previously defined, are:

1. The area between U equals 0 and U equals the AQL $\{p_l(\cdot)\}$.

2. The area between U equals the AQL and U equals CL, the control limit $\{p_h(\cdot)\}$.

3. The area beyond U equals the (upper) control limit $\{p_d(\cdot)\}$.

The transition probabilities correspond to likelihoods of observing values of the control statistic (U) in different zones on the control chart. The physical characteristics of zones on the chart are dependent on the selection of AQL, a control limit specification, a set of sampling intensities when on sampling inspection, and a subgroup size when on detailing inspection. Once the physical characteristics of the zones are quantified in terms of critical numbers of defects, transition probabilities can be easily calculated for any assumed incoming defect rate based on the Poisson distribution.

Calculation of Transition Probabilities

For any value of the process parameter, θ, the transition probabilities are given by:

(5)
$$P_l(x) = \sum_{j=0}^{N(x)} \frac{(\theta \cdot h_x)^j}{j!} e^{-\theta h_x}$$

(6)
$$p_h(x) = \sum_{j=N(x)+1}^{\text{DMAX}^{(x)}} \frac{(\theta \cdot h_x)^j}{j!} e^{-\theta h_x}$$

and

(7)
$$p_d(x) = 1 - p_l(x) - p_h(x)$$

for $x = n, r, t$, and where $N(x)$ and $\mathrm{DMAX}(x)$ are defined as follows:

(8)
$$N(x) = \max_k \left\{ k \,\middle|\, \frac{\mathrm{AQL} \cdot h_x}{100} \geqslant k;\ k = 0, 1, \ldots \right\}$$

(9)
$$\mathrm{DMAX}(x) = \operatorname*{Max}_d \left\{ d \,\middle|\, \sum_{i=0}^{d} \frac{\left[\left(\frac{\mathrm{AQL}}{100} \right) \cdot h_x \right]^i}{i!} \exp \left(- \frac{\mathrm{AQL}}{100} h_x \right) \leqslant .99;\ d = 0, 1, \ldots, \right\}$$

where $x = n, r, t$. Thus, when the number of defects found is less than or equal to $N(x)$, the point plot will always be in the control chart zone between 0 and the AQL. In terms of the physical characteristics of the control chart:

(10)
$$\mathrm{CL}_x = \left[\frac{\mathrm{DMAX}(x) + 0.5}{h_x} \right] \cdot 100 \text{ (for U in defects/100MH)}$$

The factor 0.5 is included so that points will fall above or below the CL and never on it.

It should be noted that the values of the control limit, CL_x, are based on standard sample sizes in hours and on standard values of the central line, the AQL. Thus the control chart test is whether or not the process, or shop, is operating at or below the designated AQL.

It is quite possible for a shop to be operating in a state of statistical control, but at a quality level above the AQL, with the result that values of U will fall above CL_x with high probability and thus signal a lack of control. Correspondingly it is possible for a shop to be technically out of statistical control at quality levels significantly below the AQL with the result of no signal of lack of control. This simply illustrates the difference between a standard control chart based on a "best estimate" of the actual process level and an Acceptance Control Chart based on a standard or aimed-at value of the control statistic.

3YSTEM DEVELOPMENT

Initial System Characteristics Model

A set of FORTRAN programs was developed which included calculation of state transition probabilities, steady state Markov chain probabilities, the time-weighted system probabilities ($P'(x)$, $x = R, N, T, D$), and the AFI and AOQ values for appropriate ranges of the defect rate parameter, θ. A flow diagram of the operation of this routine is shown in Figure 4. A sample of the graphical output, using a CALCOM plotter, is shown in Figure 5.

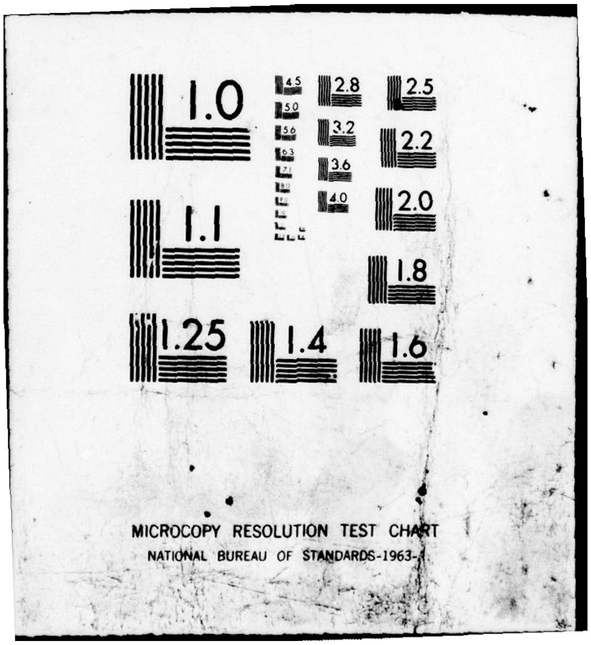
Initially, the output of this program set, coupled with printed data listing values of the AFI at θ equals the AQL, the value of θ at which the AOQ function reaches its maximum, the sample hours under Normal inspection (h_n), and the highest count of defects that lies within

AD-A075 939 OFFICE OF NAVAL RESEARCH ARLINGTON VA F/G 15/5
NAVAL RESEARCH LOGISTICS QUARTERLY. VOLUME 26, NUMBER 3.(U)
SEP 79

UNCLASSIFIED NL

2 OF 2
AD
A075939

END
DATE
FILMED
11-79
DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

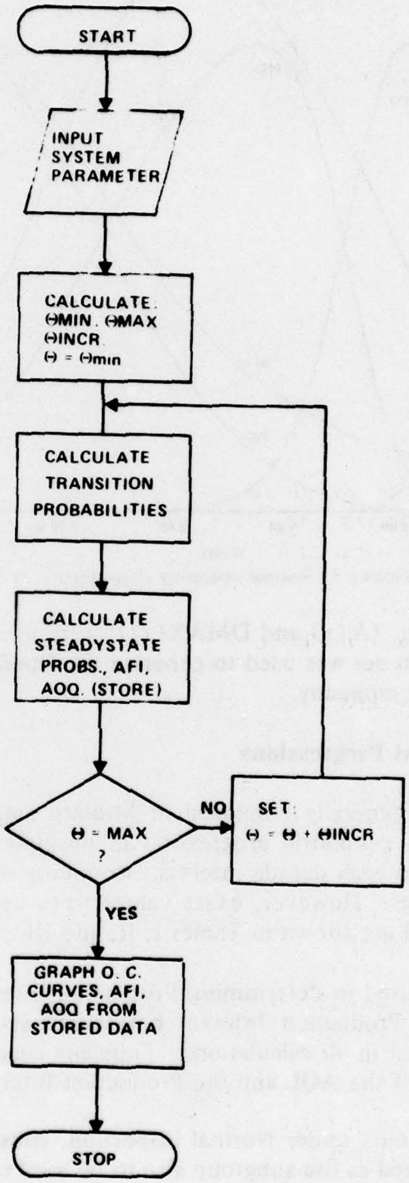

START
INPUT SYSTEM PARAMETER
CALCULATE: ΘMIN. ΘMAX ΘINCR. Θ = Θmin
CALCULATE TRANSITION PROBABILITIES
CALCULATE STEADYSTATE PROBS., AFI, AOQ. (STORE)
Θ = MAX ?
NO
SET: Θ = Θ + ΘINCR
YES
GRAPH O. C. CURVES, AFI, AOQ FROM STORED DATA
STOP

FIGURE 4. Operating characteristics program flow chart

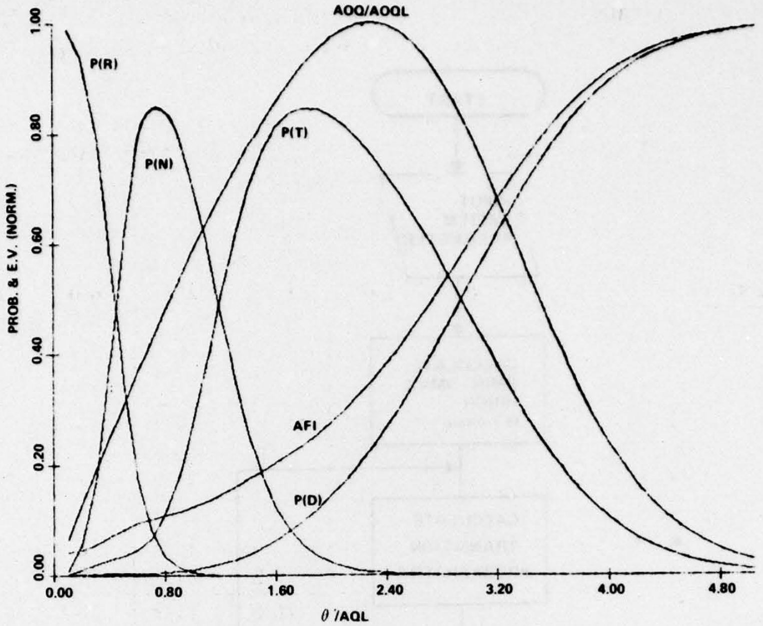


FIGURE 5. Sample operating characteristic curves

each zone of the control chart, ($N(x)$ and DMAX(x)), were used to test input system parameters. Ultimately, this program set was used to generate the Operating Characteristic Curves for the scheme developed for the company.

AQL and Production Interval Progressions

Following the practice generally employed in Military Standards, preferred AQL values were established based on a geometric progression in multiples of $10^{0.2}$. This provides five geometrically spaced values in each decade interval. Rounding of the exact geometric values is indicated in all table headings. However, exact values were used in all calculations. Sample Tables for Inspection Level II are shown in Tables I, II, and III.

The same pattern was used in determining Production Intervals, for which H varies from 100 to 10,000 man-hours. Production Interval ranges are set at the geometric mid-points between the exact values used in all calculations. Thus any specific sampling system chosen is only precise when the value of the AQL and the Production Interval are exact powers of $10^{0.2}$.

Since h_n, the sample hours under Normal inspection, constitutes the base from which h_r and h_t are derived, h_n was fixed as the subgroup size to be used on detailing inspection.

Fixing AQL-AOQL Relationship

For the scheme to be fully operational in a wide range of circumstances, it was decided that there should be a fixed relationship between designated AQL values and their related AOQL values. If values of the two are required to be identical, sample sizes, in general, will be driven unreasonably high. Therefore, it was decided to designate a relative one-step difference wherein all sampling systems for a specified AQL have a AOQL equal to $10^{0.2}$ AQL. This system is designated Inspection Level II, and was recommended to the company for general use.

Both tighter (AOQL equal to AQL) and looser (AOQL equal to $10^{0.4}$ AQL) systems are provided. These are designated Inspection Levels III and I, respectively. Tables I, II, and III give the Sample Hours, Control Limits, and AFI values evaluated at the AQL, respectively, for Inspection Level II.

TABLE I. *Sample Hours in Man-Hours of Production Under Reduced (H_r), Normal (H_n), and Tightened (H_t) Inspection. Inspection Level II.*

Production Interval (Man-Hours)		Acceptable Quality Level (AQL) in Defects per 100 Man-Hours							
		0.40	0.65	1.0	1.5	2.5	4.0	6.5	10.0
0-125 (1)	Hr						5.0	4.2	3.0
	Hn	↓	↓	↓	↓	↓	12.5	10.4	7.5
	Ht						19.8	16.6	11.9
126-200 (2)	Hr					7.9	6.6	4.7	4.3
	Hn	↓	↓	↓	↓	19.8	16.6	11.9	10.8
	Ht					31.4	26.2	18.9	17.1
201-315 (3)	Hr				12.5	10.4	7.5	6.8	6.2
	Hn	↓	↓	↓	31.4	26.2	18.9	17.1	15.6
	Ht				49.8	41.6	29.9	27.1	24.7
316-500 (4)	Hr			19.8	16.6	11.9	10.8	9.8	8.0
	Hn	↓	↓	49.8	41.6	29.9	27.1	24.7	20.0
	Ht			79.0	65.9	47.4	42.9	39.2	31.7
501-800 (5)	Hr		31.4	26.2	18.9	17.1	15.6	12.6	9.9
	Hn	↓	79.0	65.9	47.4	42.9	39.2	31.7	24.9
	Ht		125	104	75.1	68.0	62.1	50.2	39.4
801-1,250 (6)	Hr	49.8	41.6	29.9	27.1	24.7	20.0	15.7	11.2
	Hn	125	104	75.1	68.0	62.1	50.2	39.4	28.0
	Ht	198	166	119	108	98.4	79.5	62.4	44.4
1,251-2,000 (7)	Hr	65.9	47.4	42.9	39.2	31.7	24.9	17.7	14.8
	Hn	166	119	108	98.4	79.5	62.4	44.4	37.2
	Ht	262	189	171	156	126	99.0	70.4	58.9
2,001-3,160 (8)	Hr	75.1	68.0	62.1	50.2	39.4	28.0	23.5	19.2
	Hn	189	171	156	126	99.0	70.4	58.9	48.3
	Ht	299	271	247	200	157	112	93.4	76.5
3,161-5,000 (9)	Hr	108	98.4	79.5	62.4	44.4	37.2	30.5	21.9
	Hn	271	247	200	157	112	93.4	76.5	55.1
	Ht	430	392	317	249	177	148	121	87.3
5,001-8,000 (10)	Hr	156	126	99.0	70.4	58.9	48.3	34.8	25.2
	Hn	392	317	249	177	148	121	84.3	63.2
	Ht	621	502	394	280	235	192	138	100
Over 8,000 (11)	Hr	200	157	112	93.4	76.5	55.1	39.9	28.5
	Hn	502	394	280	235	192	138	100	71.5
	Ht	795	624	444	372	305	219	159	113
		0.65	1.0	1.5	2.5	4.0	6.5	10.0	15.0
		Average Outgoing Quality Limit (AOQL) in Defects per 100 Man-Hours							

↓ Proceed in direction of arrow until first plan is encountered.

TABLE II — *Control Limits under Reduced (CLr), Normal (CLn) and Tightened (CLt) Inspection in Defects per 100 Man-Hours. Inspection Level II.*

Production Interval (Man-Hours)		Acceptable Quality Level (AQL) in Defects per 100 Man-Hours							
		0.40	0.65	1.0	1.5	2.5	4.0	6.5	10.0
0-125	CLr						30.0	36.0	50.0
	CLn	↓	↓	↓	↓	↓	20.0	24.0	33.0
	CLt						12.5	21.0	30.0
126-200	CLr					19.0	23.0	32.0	35.0
	CLn	↓	↓	↓	↓	12.5	15.0	21.0	32.0
	CLt					8.0	13.5	18.5	26.0
201-315	CLr				12.0	14.5	20.0	22.0	40.0
	CLn	↓	↓	↓	8.0	9.5	13.5	20.5	29.0
	CLt				5.0	8.5	12.0	16.5	26.0
316-500	CLr			7.5	9.0	12.5	14.0	25.5	31.0
	CLn	↓	↓	5.0	6.0	8.5	13.0	18.0	27.0
	CLt			3.0	5.5	7.5	10.5	16.5	24.0
501-800	CLr		4.5	5.5	8.0	8.5	16.0	20.0	35.0
	CLn	↓	3.0	4.0	5.5	8.0	11.5	17.0	26.0
	CLt		2.0	3.5	4.5	6.5	10.5	15.0	22.0
801-1,250	CLr	3.0	3.5	5.0	5.5	10.0	12.5	22.0	31.0
	CLn	2.0	2.5	3.5	5.0	7.0	11.0	16.5	23.0
	CLt	1.3	2.0	3.0	4.0	6.5	9.5	13.5	22.0
1,251-2,000	CLr	2.5	3.0	3.5	6.5	8.0	14.0	20.0	30.0
	CLn	1.5	2.0	3.5	4.5	7.0	10.5	14.5	23.0
	CLt	1.3	1.8	2.5	4.0	6.0	8.5	13.5	20.0
2,001-3,160	CLr	2.0	2.5	4.0	5.0	9.0	12.5	19.0	29.0
	CLn	1.5	2.0	3.0	4.5	6.5	9.0	14.5	22.0
	CLt	1.2	1.6	2.5	3.8	5.5	8.5	12.5	19.0
3,161-5,000	CLr	1.5	2.5	3.0	5.5	8.0	12.0	18.0	25.0
	CLn	1.3	1.8	2.8	4.0	5.8	9.0	13.5	21.0
	CLt	1.0	1.6	2.4	3.5	5.4	7.8	12.0	18.0
5,001-8,000	CLr	1.6	2.0	3.5	5.0	7.5	11.5	16.0	26.0
	CLn	1.2	1.7	2.5	3.6	5.8	8.5	13.0	20.0
	CLt	1.0	1.5	2.2	3.4	5.0	7.5	11.0	17.5
Over 8,000	CLr	1.2	2.2	3.2	4.5	7.0	10.0	16.0	23.0
	CLn	1.1	1.6	2.4	3.5	5.5	8.5	12.5	19.0
	CLt	1.0	1.4	2.2	3.1	4.8	7.0	11.0	17.0
		0.65	1.0	1.5	2.5	4.0	6.5	10.0	15.0
		Average Outgoing Quality Limit (AOQL) in Defects per 100 Man-Hours							

↓ Proceed in direction of arrow until first plan is encountered.

TABLE III – *Average Fraction Inspected (AFI) at the Acceptable Quality Level (AQL), in Percent of Shop Man-Hours Inspection Level II.*

Production Interval	Acceptable Quality Level (AQL) in Defects per 100 Man-Hours							
(Man-Hours)	0.40	0.65	1.0	1.5	2.5	4.0	6.5	10.0
0-125	↓	↓	↓	↓	↓	11.59	10.75	9.11
126-200	↓	↓	↓	↓	11.59	10.75	9.11	8.56
201-315	↓	↓	↓	11.59	10.75	9.11	8.56	6.47
316-500	↓	↓	11.59	10.75	9.11	8.56	6.47	5.82
501-800	↓	11.59	10.75	9.11	8.56	6.47	5.82	4.25
801-1,250	11.59	10.75	9.11	8.56	6.47	5.82	4.25	3.14
1,251-2,000	10.75	9.11	8.56	6.47	5.82	4.25	3.14	2.44
2,001-3,160	9.11	8.56	6.47	5.82	4.25	3.14	2.44	2.00
3,161-5,000	8.56	6.47	5.82	4.25	3.14	2.44	2.00	1.40
5,001-8,000	6.47	5.82	4.25	3.14	2.44	2.00	1.40	0.88
Over 8,000	5.82	4.25	3.14	2.44	2.00	1.40	0.88	0.72
	0.65	1.0	1.5	2.5	4.0	6.5	10.0	15.0
	Average Outgoing Quality Limit (AOQL) in Defects per 100 Man-Hours							

↓Proceed in direction of arrow until first plan is encountered.

System Optimization

With the probability values associated with the CL, the general system design, and the relationship among the sample hours under various levels of inspection fixed, it remained to determine the sample hours, h_n, and values of the k_i's, the run lengths for the switching rules, in order to:

1. Meet fixed values of AOQL, and
2. Minimize the AFI if the process is operating at (or below) the AQL.

Relationships Among Sample Hours

Initial study showed that system characteristics were more sensitive to average sample hours and to selected run lengths (the values of the k_i's in Figure 1) than to the relationship between the sample hours, h_r, h_n, and h_t, specified for a given system. This relationship afforded a great deal of flexibility in the choice of relationship between h_r, h_n, and h_t.

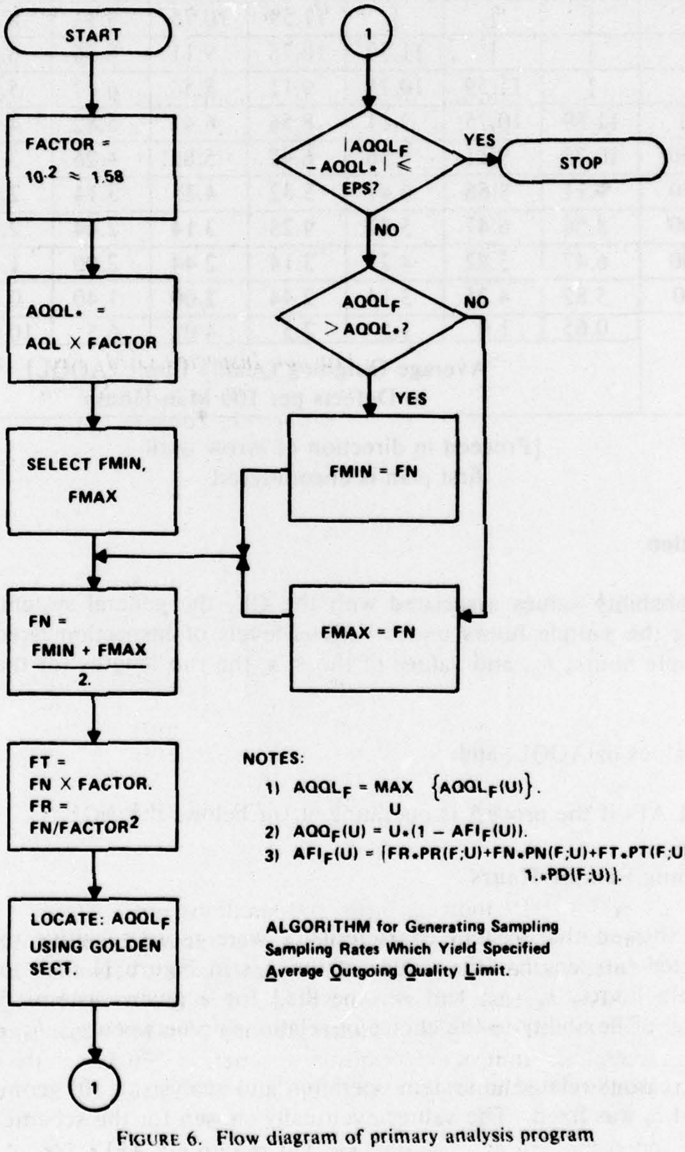
For practical reasons related to system operation and analysis, a 10^x geometric relationship between h_r, h_n, and h_t was fixed. The values eventually chosen for the scheme were:

$$h_r = 10^{-0.4}\, h_n, \text{ and}$$

$$h_t = 10^{0.2}\, h_n.$$

The net result was that, along any secondary diagonal of the system (i.e., moving diagonally from lower left to upper right in a table), the parameter of the Poisson distribution function, θh_n, remains constant. In addition, the same Poisson parameters apply across three diagonals, one for h_r, h_n, and h_t. Among other technical advantages relating to system analysis and (attempted) optimization, this structure permitted a substantial reduction in the number of Operating Characteristic Curves required to fully describe an Inspection Level system of plans.[2]

A FORTRAN Program was developed which permitted the input of a number of combinations of the k_i's, and iteratively solves for the values of h_n (and thus h_r and h_t) yielding the required AOQL. A flow diagram of this program is shown in Figure 6.


START
FACTOR = $10^{.2} \approx 1.58$
AQQL* = AQL × FACTOR
SELECT FMIN, FMAX
FN = $\frac{FMIN + FMAX}{2.}$
FT = FN × FACTOR. FR = FN/FACTOR^2
LOCATE: AQQL$_F$ USING GOLDEN SECT.
1
1
$|AQQL_F - AQQL*| \leq$ EPS?
YES
STOP
NO
$AQQL_F > AQQL*$?
NO
YES
FMIN = FN
FMAX = FN

NOTES:

1) $AQQL_F = \underset{U}{MAX} \{AQQL_F(U)\}$.

2) $AOQ_F(U) = U*(1 - AFI_F(U))$.

3) $AFI_F(U) = [FR*PR(F;U)+FN*PN(F;U)+FT*PT(F;U)+1.*PD(F;U)]$.

ALGORITHM for Generating Sampling Sampling Rates Which Yield Specified Average Outgoing Quality Limit.

FIGURE 6. Flow diagram of primary analysis program

[2] There are 73 sampling systems specified for Inspection Level II containing three separate sampling plans each. However, only 13 sets of Operating Characteristics Curves are necessary to fully describe the system.

Just as the parameter of the Poisson distribution function remains constant along any secondary diagonal in the scheme, the product of the AQL times the Production Interval, H, also remains constant. It is therefore convenient for analytical purposes to work with the ratios of the sample hours to the Production Intervals, $f_x = h_x/H$, rather than with h_x directly. The flow diagram reflects this fact.

In each specification of an f_n, it is necessary to search the AOQ function to locate the value of θ at which the function reaches its maximum. This is accomplished by the method of Golden Sections [10, p. 537]. The AOQ function is then evaluated at this point to obtain the AOQL. The AOQL found is then tested against the required value. The value of f_n is then adjusted either upward to downward using a bisection search routine until the required AOQL is obtained. Once this procedure has been carried through, the resulting values of f_n, f_r, and f_t apply to all systems along the applicable secondary diagonal.

The Analysis Program was run on a number of combinations of switching rules (values of the k_i's) and geometric relationships between f_n and f_r and f_t. Values of the AFI evaluated at the AQL were then tabulated in a form similar to that illustrated in Table IV for each secondary diagonal. Results for secondary diagonals 6 and 10 are illustrated here because the region between these diagonals includes almost all plans currently in use by the company. Secondary diagonals may be identified in Table I by the number in parentheses beneath the Production Interval designation in the first column.

TABLE IV—*Values of Average Fraction Inspected (AFI) Evaluated at the AQL for Selected Combinations of Sampling Systems for Two Secondary Diagonals of Table I.*

					Diagonal No. 6			Diagonal No. 10		
		f_r			$10^{-0.2} f_n$	$10^{-0.4} f_n$	$10^{-0.4} f_n$	$10^{-0.2} f_n$	$10^{-0.4} f_n$	$10^{-0.4} f_n$
		f_t			$10^{0.2} f_n$	$10^{0.2} f_n$	$10^{0.4} f_n$	$10^{0.2} f_n$	$10^{0.2} f_n$	$10^{0.4} f_n$
k_1	k_2	k_3	k_4	k_5						
7	3	7	3	5	X	0.132	deg.	X	0.073	0.072
5	5	7	3	5	0.152	0.118	deg.	0.073	0.069	0.064
5	5	7	3	5	0.152	0.116	deg.	0.069	0.066	0.055
5	5	5	5	5	X	0.171	deg.	X	0.081	0.093
5	5	5	3	5	0.154	0.118	deg.	0.071	0.066	0.055
5	5	5	3	3	0.187	0.159	0.130	0.068	0.082	0.094
5	3	5	3	5	X	0.120	deg.	X	0.071	0.066
5	3	5	3	3	0.188	0.134	0.163	0.076	0.098	0.086

The "X" entries in Table IV indicate those combinations not analyzed. The "deg" entries indicate degenerate plans for that combination along that secondary diagonal. Any plan must allow for at least one defect to lie in the zone of the control chart between the AQL central line and the control limit. Also, at least 2 defects are required in the zone above the Control Limit if the switching rules are to work properly. Therefore, any plan for which one defect results in a point plot above the control limit is by definition degenerate. Such was the case along diagonal 6 for most plans under the column $f_r = 10^{-0.4} f_n$, $f_t = 10^{0.4} f_n$.

It should be added that no combination yielded minimum AFI's for all diagonals. However, those under the column $f_r = 10^{-0.4} f_n$, $f_t = 10^{0.2} f_n$ and for switching rules combination (5, 5, 7, 3, 5) yielded the minimum most frequently and thus became the switching rule set chosen.

Implementation of the System

The company was provided with a procedural document contai-ing a complete set of tables similar to Tables I, II, and III, Operating Characteristic Curves similar to Figure 5, and directions on use of the system. While they were provided with the mathematical formulation of the system, their interest was in the tabulated results which allowed them to determine when they should be on detailing inspection and when on sampling.

Because of the similarity of the system design to the procedural aspects of MIL-STD-105D, there was little difficulty in making the transition to the new system. After six months of operation in more than sixty shops, the new system has produced positive results. The trend has been steadily downward in defects found in items out in the field. Also, the new system has made it possible to respond to changes in quality level much quicker than the previous system because information is processed weekly in the shop rather than monthly by a remote computer center.

BIBLIOGRAPHY

[1] Dodge, H. F., and H. G. Romig, *Sampling Inspection Tables, Single and Double Sampling*, 2nd Ed. (John Wiley & Sons, Inc., New York, 1959).

[2] Duncan, A. J., "The Economic Design of $\bar{X}$ Charts Used to Maintain Current Control of a Process," Journal of the American Statistical Association, Vol. 51, 1956.

[3] Freund, R. A., "Acceptance Control Charts," Industrial Quality Control, Oct., 1957.

[4] Grant, E. L., and R. S. Leavenworth, *Statistical Quality Control*, 4th ed., (McGraw-Hill Book Co., New York, 1972).

[5] Hill, I. D., "Sampling Inspection and Defense Specification DER-131," Journal of Royal Statistical Society, Series A, 125, Part 1, pp. 31-73 (1962).

[6] *MIL-STD-105D*, "Sampling Procedures and Tables for Inspection by Attributes," U. S. Government Printing Office, Washington, D. C. (1963).

[7] Montgomery, D. C., R. G. Heikes, and J. F. Mance, "Economic Design of Fraction Defective Control Charts," Management Science, Vol. 21, No. 11 (1975).

[8] Pyke, R., and R. Schaufele, "Limit Theorems for Markov Renewal Processes," Annals of Mathematical Statistics, pp. 1746-1764 (1964).

[9] Stephens, K. S., and K. E. Larson, "An evaluation of the *MIL-STD-105D* System of Sampling Plans," Industrial Quality Control (Jan. 1957).

[10] Wagner, H. M., *Principles of Operations Research*, 2nd ed., (Prentice-Hall, Inc., Englewood Cliffs, N. J., 1975).

Appendix A
THEORETICAL BASIS OF MODEL

Because of the nature of the sampling system, it is not necessary to have explicit functions for each of the state probabilities but only to have them in terms of the combined states under Normal (N), Reduced (R), Tightened (T), and 100 percent Detailing (D) inspection levels. For each of the sampling states, a designated sample size is drawn from a total of H hours, a Production Interval. However, when a transition is made to detailing, the time spent in each detailing state upon a transition into that state is h_n hours where $h_n = f_n H$.

Consider a Markov Renewal Process with transition matrix $P = [p_{ij}]$ as shown in Figure 3 and interarrival (or sojourn) times x_{ij}. Suppose x_{ij} denotes the length of time during which the system is in state i and going to state j, and denote the distribution function of x_{ij} by $F_{ij}(\cdot)$. For the model under consideration,

$$x_{ij} = \begin{cases} H \text{ if } i \in R \cup N \cup T \\ h_n \text{ if } i \in D \end{cases}$$

where R denotes the set of states for reduced inspection, N for normal inspection, etc. Note that $F_{ij}(\cdot)$ only depends on i. Let

(A.1) $$\underline{\pi} = (\pi_0, \pi_1, \ldots, \pi_k)$$

denote the stationary probability vector for P, and let

$$\sum_{i \in R} \pi_i = P(R)$$

$$\sum_{i \in N} \pi_i = P(N)$$

$$\sum_{i \in T} \pi_i = P(T)$$

$$\sum_{i \in D} \pi_i = P(D).$$

$P(\cdot)$ are, therefore, the steady-state probabilities of being in R, N, T, and D inspection resulting from solution of the Markov chain.

Consider any real-valued functions $f(i, j, x)$ and let J_n denote the state of the system at the n^{th} transition. Define

(A.2) $$w_f(t) = \sum_{n=1}^{N_t} f(J_{n-1}, J_n, X_n)$$

where N_t is the number of transitions in $(0, t]$ and X_n is the sojourn time between J_{n-1} and J_n.

To investigate the fraction inspected, we are interested in the specific function

(A.3) $$f(i, j, x) = \begin{cases} h_r \text{ if } i \in R \\ h_n \text{ if } i \in N \cup D. \\ h_t \text{ if } i \in T \end{cases}$$

Now, $w_f(t)/t$ is approximately the fraction inspected out of t standard man-hours of labor.

The limiting properties of $w_f(t)$ are found by looking at the return times of a specific state, say state 0, and considering

$$y_n = w_f(T_{0,\, n+1}) - w_f(T_{0,n})$$

where $t_{0,n}$ is the time of the n^{th} return to 0.

The following notation is required:

$$_j{}^M{}_{ki} = E[\text{number of visits to state } i \text{ until entry } j, \text{ starting from } k]$$

$$_0{}^M{}_{0i} \equiv m_i = (\pi_i/\pi_0)$$

$$\xi_{ij}(f) = \int_{-\infty}^{\infty} f(i, j, x) p_{ij} dF_{ij}(x)$$

$$\xi_i(f) = \sum_j \xi_{ij}(f)$$

$$\xi_{ij}^{(2)}(f) = \int_{-\infty}^{\infty} f^2(i, j, x)\, p_{ij} dF_{ij}(x)$$

$$\xi_i^{(2)}(f) = \sum_j \xi_{ij}^{(2)}(f)$$

Certain basic results then follow from [8]. If $\xi_i(|f|) \leqslant \infty$, then

$$E(Y_n) = \sum_i m_i \xi_i(f)$$

Letting μ_{00} denote the mean recurrence time of state 0 and $\eta_i = E(x_{ij})$, it follows that:

$$\mu_{00} = \sum_i \eta_i m_i$$

Under certain general conditions, as $t \to \infty$,

$$\frac{w_f^{(t)}}{t} \xrightarrow{a.s.} \frac{E(Y_n)}{\mu_{00}}$$

For $f(i, j, x)$ as in equation A.3,

$$\sum_i \pi_i \xi_i(f) = h_r P(R) + h_n P(N) + h_t P(T) + h_n P(D) = \pi_0 E(Y_n)$$

and

$$\sum_i \pi_i \eta_i = h_n P(D) + H(1 - P(D)) = \pi_0 \mu_{00}.$$

Therefore, letting FI denote "fraction inspected," we have

$$(A.4) \qquad FI \xrightarrow{a.s.} \frac{h_r P(R) + h_t P(T) + h_n [P(N) + P(D)]}{h_n P(D) + H(1 - P(D))} = AFI$$

This justifies the results given for the AFI in the section titled "Formulation of Operating Characteristics of the System Model," and leading to equation 5. Other functions can be defined to give further properties of the system. For example, let

$$f_r(i, j, x) = \begin{cases} H, & \text{if } i \in R \\ 0, & \text{otherwise} \end{cases}$$

The the fraction of time spent in reduced inspection is given by

$$\frac{\sum_{n=1}^{N_t} f_r(J_{n-1}, J_n, X_n)}{t} \xrightarrow{a.s.} \frac{HP(R)}{h_n P(D) + H[1 - P(D)]}$$

with similar results, except for state D, for the other three inspection states. For state D,

$$f_d(i, j, x) = \begin{cases} h_n, & \text{if } i \in D \\ 0, & \text{otherwise} \end{cases}$$

and

$$\frac{\sum_{n=1}^{N_t} f_d(J_{n-1}, J_n, X_n)}{t} \xrightarrow{a.s.} \frac{h_n P(D)}{h_n P(D) + H[1 - (D)]}$$

These results are useful in describing the operating characteristics, and corresponding Operating Characteristic (O.C.) Curves, for the sampling systems. The formulas are developed in the preceding section of 'he paper wherein $P(\cdot)$ denotes, as above, the steady state probabilities of R, N, T, and D inspection resulting from solution of the Markov chain. $P'(\cdot)$ denotes the time-weighted probabilities of the respective states. Thus, for example:

(A.5)
$$P'(R) = \frac{HP(R)}{h_n P(D) + H[1 - P(D)]}$$

APPENDIX B
GENERAL SOLUTION OF THE MARKOV CHAIN

Steady-State System Equations

The equations describing the steady state probabilities for the general system are as follows where:

$P\ (X)$ = non time-weighted probability of being in state X.
R_i, N_i, N', T_i, and D_i = sub-states of Reduced (R), Normal (N), Tightened (T), and Detailing (D) states, respectively.

(B.1a) $P(R_i) = p_h(r)P(R_{i-1}) = p_h(r)^{i-1}P(R_1)$ for $i = 2, 3, \ldots, k_2$

(B.1b) $P(R_1) = p_l(r)\,[\sum_{i=1}^{k_2} P(R_i)] + p_l(n)P(N_{k_1})$

(B.1c) $P(R) = \sum_{i=1}^{k_2} P(R_i) = \left[\frac{1 - p_h(r)^{k_2}}{1 - p_h(r)}\right] P(R_1)$

(B.2a) $P(T_i) = p_l(t)P(T_{i-1}) = p_l(t)^{i-1}P(T_1)$ for $i = 2, 3, \ldots, k_4$

(B.2b) $P(T_1) = p_h(t)\ [\sum_{i=1}^{k_4} P(T_i)] + p_h(n)P(N^{k_3}) + [p_l(n) + p_h(n)]\ P(D_{k_5})$

(B.2c) $P(T) = \sum_{i=1}^{k_4} P(T_i) = \left[\frac{1 - p_l(t)^{k_4}}{1 - p_l(t)}\right] P(T_1)$

(B.3a) $P(D_i) = [p_l(n) + p_h(n)]\ P(D_{i-1}) = [p_l(n) + p_h(n)]^{i-1}P(D_1)$
for $i = 2, 3, \ldots, k_5$

(B.3b)
$$\begin{aligned} P(D_1) = {} & [1 - p_l(r) - p_h(r)]\ [\sum_{i=1}^{k_2} P(R_i)] \\ & + [1 - p_l(n) - p_h(n)]\ [\sum_{i=1}^{k_1} P(N_i) + \sum_{i=2}^{k_3} P(N^i)] \\ & + [1 - p_l(t) - p_h(t)]\ [\sum_{i=1}^{k_4} P(T_i)] \\ & + [1 - p(d)]\ [\sum_{i=1}^{k_5} P(D_i)] \end{aligned}$$

(B.3c)
$$P(D) = \sum_{i=1}^{k_5} P(D_i) = \left[\frac{1 - p(d)^{k_5}}{1 - p(d)} \right] P(D_1)$$

(B.4a)
$$P(N_i) = p_l(n) P(N_{i-1}) = p_l(n)^{i-2} P(N_2) \text{ for } i = 3, 4, \ldots, k_1$$

(B.4b)
$$P(N_2) = p_l(n) \, [P(N_1) + \sum_{i=2}^{k_3} P(N^i)]$$

(B.4c)
$$\sum_{i=2}^{k_1} P(N_i) = \left[\frac{1 - p_l(n)^{k_1 - 1}}{1 - p_l(n)} \right] P(N_2)$$

(B.5)
$$P(N_1) = p_h(r) P(R_{k_2}) + p_l(t) P(T_{k_4})$$

(B.6a)
$$P(N^i) = p_h(n) P(N^{i-1}) = p_h(n)^{i-2} P(N^2) \text{ for } i = 3, 4, \ldots, k_3$$

(B.6b)
$$P(N^2) = p_h(n) \, [P(N_1) + \sum_{i=2}^{k_1} P(N_i)]$$

(B.6c)
$$\sum_{i=2}^{k_3} P(N^i) = \left[\frac{1 - p_h(n)^{k_3 - 1}}{1 - p_h(n)} \right] P(N^2)$$

In solving the system of equations,

$$p(d) = p_l(n) + p_h(n)$$

since the sample subgroup size for detailing is the same as that used under the normal level of sampling inspection. The equation for $P(D_1)$ is replaced by the system requirement that all steady state probabilities must sum to 1. Thus $P(D_1)$, equation (B.3b), is replaced by:

(B.7a)
$$P(R) + P(T) + P(N) + P(D) = 1$$

where:

(B.7b)
$$P(N) = P(N_1) + \sum_{i=2}^{k_1} P(N_i) + \sum_{i=2}^{k_3} P(N^i).$$

General Solution of System Steady-State Probabilities

An outline of the method used to obtain the non time-weighted steady state probabilities for the major system states R, N, T, and D, is given in this section.

From equations B.1c, B.2c, B.3c, B.4c, B.5, and B.6c, it is clear that the steady state probabilities for the major system states can be written in terms of the steady state probabilities of six of the substates, regardless of the order of the system. The required substates are: $P(D_1)$, $P(T_1)$, $P(R_1)$, $P(N_1)$, $P(N_2)$, and $P(N^2)$.

For notational simplicity, define:

(B.8) $S_l(n) = \sum_{i=1}^{k_1} p_l(n)^{i-1} = [1 - p_l(n)^{k_1}]/[1 - p_l(n)]$

Similarly, $S_h(n)$, $S(r)$, $S(t)$ and $S(d)$ are defined as in (B.8) with $p_l(n)$ replaced by $p_h(n)$, $p_h(r)$, $p_l(t)$ and $p(d)$, respectively, and with k_1 replaced by k_3, k_2, k_4, and k_5, respectively. Also let,

(B.9) $g = [1 - p_l(r)S(r)]/[p_l(n)^{k_1 - 1}]$.

Using equations B.1 through B.7 and the notational forms of B.8 and B.9, the following six equations are obtained:

(B.10) $p_l(t)^{k_4} P(T_1) + p_h(r)^{k_2} P(R_1) - P(N_1) = 0$

(B.11) $p_l(n) \cdot P(N_1) + \dfrac{p_l(n)}{p_h(n)} [S_h(n) - 1] P(N^2) - P(N_2) = 0$

(B.12) $p_h(n) \cdot P(N_1) + \dfrac{p_h(n)}{p_l(n)} [S_l(n) - 1] P(N_2) - P(N^2) = 0$

(B.13) $[p_h(t)S(t) - 1] P(T_1) + p_h(n)^{k_3 - 1} P(N^2) + p(d)^{k_5} P(D_1) = 0$

(B.14) $[p_l(r)S(r) - 1] P(R_1) + p_l(n)^{k_1 - 1} P(N_2) = 0$

(B.15) $S(r)P(R_1) + S(t)P(T_1) + S(d)P(D_1) + P(N_1) + \dfrac{S_l(n) - 1}{p_l(n)} P(N_2) +$

$\dfrac{S_h(n) - 1}{p_h(n)} P(N^2) = 1$

Using equation B.14, noting that $P(N_2) = gP(R_1)$, and making appropriate substitutions, a system of five equations in five unknowns can be derived which can be solved by any suitable method.

A Cramer's Rule solution was applied to the preceeding equations to obtain the following results for the major system states:

(B.16) $P(D) = S(d)P(D_1) = Z_1/C$

(B.17) $P(T) = S(t)P(T_1) = Z_2/C$

(B.18) $P(R) = S(r)P(R_1) = Z_3/C$

(B.19) $P(N) = 1 - P(D) - P(T) - P(R)$,

where:

$$Z_1 = S(d)\{p_h(r)^{k_2} p_l(n)S_h(n) [1 - p_h(t)S(t)] - g[1 - p_h(t)S(t)]$$

$$[S_h(n) + S_l(n) - S_h(n)S_l(n)] + gp_h(n)^{k_3} p_l(t)^{k_4} S_l(n)\}$$

$$Z_2 = S(t)p(d)^{k_5} \{p_h(r)^{k_2} p_l(n) S_h(n) - g[S_h(n) + S_l(n) - S_h(n)S_l(n)]\}$$

$$Z_3 = S(r)\{-p(d)^{k_5} p_l(t)^{k_4} p_l(n)S_h(n)\}$$

and,

$$C = Z_1 + Z_2 + Z_3 - g \cdot p(d)^{k_5} p_l(t)^{k_4} S_h(n)S_l(n).$$

A SINGLE SERVER QUEUE WITH ARRIVAL RATE DEPENDENT ON SERVER BREAKDOWNS

Andrew W. Shogan

University of California, Berkeley
Berkeley, California

ABSTRACT

This paper considers a single server queueing system that alternates stochastically between two states: operational and failed. When operational, the system functions as an $M/E_k/1$ queue. When the system is failed, no service takes place but customers continue to arrive according to a Poisson process; however, the arrival rate is different from that when the system is operational. The durations of the operating and failed periods are exponential with mean $1/c\alpha$ and Erlang with mean $1/c\beta$, respectively. Generating functions are used to derive the steady-state quantities L and W, both of which, when viewed as functions of c, decrease at a rate inversely proportional to c^2. The paper includes an analysis of several special and extreme cases and an application to a production-storage system.

In many queueing systems (e.g., a computer facility), the server is subject to breakdown. If the breakdowns are unpredictable in nature and the server is not replaced or repaired until a breakdown occurs, then the facility will be unable to provide uninterrupted service to its customers. In such a case, it is important to understand how the breakdowns will affect the system's level of performance.

Avi-Itzhak and Naor [1] and Gaver [3] obtained the steady state quantities L and W for an $M/G/1$ queue whose unreliable server alternates between operational periods of exponential duration and failed periods of arbitrary random duration. Although general with respect to the distributions of the service and repair times, the results have the disadvantage of requiring a Poisson arrival process with a constant rate; that is, an arrival rate *not* dependent upon whether the server is operational or failed.

Yechiali and Naor [7] and Fond and Ross [2] both considered variations of a single server exponential queueing model in which the arrival and service rates alternate between the pairs (λ_1, μ_1) and (λ_2, μ_2), with the length of time the system operates with pair (λ_i, μ_i) being exponentially distributed with rate $c\alpha_i$, $i = 1, 2$. In [7], the steady state quantity L was derived for a system with infinite queue capacity while, in [2], the steady-state proportion of customers lost was obtained for a system in which any arrival finding the server busy is lost. The special case $\mu_2 = 0$ results in models of queueing systems subject to breakdown and having Poisson arrival rates dependent upon the operation or breakdown of the server; however, all probability distributions are restricted to be exponential.

The model developed here is an important one that cannot be obtained as a special case of the models in the literature. In particular, consider a single server queueing system with the following characteristics:

(a) The system alternates between two states: the operational state and the failed state (also referred to as the repair state).

(b) When operational, the system functions as an $M/E_k/1$ queue; that is, customers arrive according to a Poisson process with rate λ, and service is according to an Erlang distribution with mean $1/\mu$ and shape parameter k.

(c) If service to a customer is interrupted by a breakdown, resumption takes place as soon as the repair period ends with no loss of service involved.

(d) Although no service takes place during the repair period, customers continue to arrive according to a Poisson process, but now having a rate of λ_1 instead of λ.

(e) The duration of operating periods is exponential with mean $1/c\alpha$ and the duration of repair periods is Erlang with mean $1/c\beta$ and shape parameter m.

This model is not a special case of [7] because it permits both the service and repair times to have Erlang distributions rather than restricting them to be exponential. Furthermore, the model is not a special case of [1] and [3] because the Poisson arrival rate is state dependent when $\lambda_1 \neq \lambda$. Such an arrival process is useful in many practical situations where the customers are aware of when the server is inoperable; in such cases, one expects to find $\lambda_1 < \lambda$ or even $\lambda_1 = 0$.

The constant c in assumption (e) controls how rapidly the system oscilates between the operational and failed states. Holding α and β constant while increasing c has the effect of keeping the steady state probabilities of being in the operational and failed states constant while increasing the frequency with which the system changes states.

The remainder of this paper is organized as follows: Section 1 analyzes the queueing system described by assumptions (a) - (e). Some special and extreme cases of the general model are considered in Section 2. Section 3 not only shows that, as functions of c, both L and W decrease at a rate inversely proportional to c^2 but also investigates the behavior of the system as $c \to \infty$. The paper concludes in Section 4 with an application to a production-storage system.

1. STEADY STATE RESULTS

Preliminary Analysis

Throughout this section, as well as Sections 2 and 4, the constant c will be assumed to equal 1. This eliminates the need to carry c along in all the derivations when it is only relevant to the analysis in Section 3.

The "method of phases" (cf. [4, p. 168]) provides a convenient means of obtaining the steady state results. It is well-known that an Erlang random variable with mean $1/\mu$ and shape parameter k is equivalent to the sum of k independent, exponentially distributed random variables each having the same mean $1/k\mu$. Hereafter, both the Erlang service times and Erlang repair times of the model will be viewed as consisting of a series of identical and independent, exponentially distributed phases.

The system can now be analyzed as a continuous time Markov process with states $\{(i, j) \mid i = 0, 1, \ldots, m$ and $j = 0, 1, 2, \ldots\}$ where $i = 0$ denotes the system is operational, $1 \leqslant i \leqslant m$ denotes the number of phases remaining in the repair process until the system becomes operational, and j denotes the number of service phases in the system (the sum of the number of phases remaining for the customer in service and k times the number of customers in the queue). The transition probabilities are stationary and satisfy the Kolmogorov differential equations. Furthermore, the steading state probabilities $\{p_{ij}\}$ exist, are independent of the initial state, and satisfy the following balance equations:

$$\text{(1a)} \qquad (\lambda_1 + m\beta) \quad p_{m0} = \alpha \; p_{00} \qquad (j = 0)$$

$$\text{(1b)} \qquad (\lambda_1 + m\beta) \quad p_{i0} = m\beta \; p_{i+1,0} \quad (1 \leqslant i \leqslant m-1) \qquad (j = 0)$$

$$\text{(1c)} \qquad (\lambda + \alpha) \quad p_{00} = m\beta \; p_{10} \quad + k\mu p_{01} \qquad (j = 0)$$

$$\text{(1d)} \qquad (\lambda_1 + m\beta) \quad p_{mj} = \alpha \; p_{0j} \; + \lambda_1 \, p_{m,j-k} \qquad (j > 0)$$

$$\text{(1e)} \qquad (\lambda_1 + m\beta) \quad p_{ij} = m\beta \; p_{i+1,j} + \lambda_1 p_{i,j-k} \quad (1 \leqslant i \leqslant m-1) \qquad (j > 0)$$

$$\text{(1f)} \qquad (\lambda + \alpha + k\mu) \quad p_{0j} = m\beta \; p_{1j} \quad + \lambda \; p_{0,j-k} + k\mu p_{0,j+1} \qquad (j > 0)$$

where a negative subscript in (1d) - (1f) indicates the term is zero. Figure 1 contains the portion of the Markov chain's state transition diagram corresponding to states (i, j) with $0 \leqslant i \leqslant m$ and $j \geqslant k$. It is clear from the figure that equations (1) can be interpreted as requiring the mean transition rates into and out of a state to be equal at steady state.

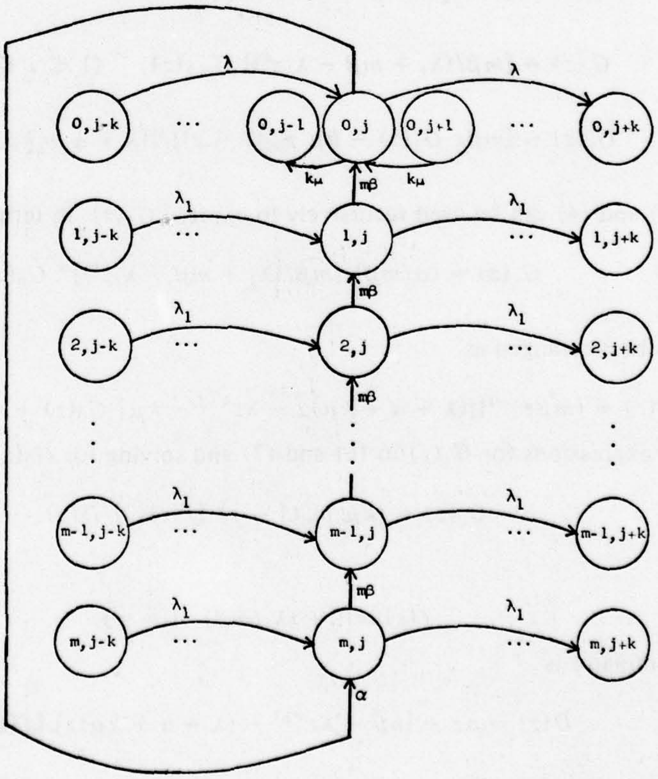


FIGURE 1.

Let $p_w = \Sigma_{j=0}^{\infty} p_{0j}$ and $p_f = \Sigma_{i=1}^{m} \Sigma_{j=0}^{\infty} p_{ij}$; that is, p_w and p_f are the steady state probabilities of the system being operational and failed, respectively. On considering the underlying two state (operational and failed) stochastic process, it is immediate that

$$p_w = \beta/(\alpha + \beta),$$

$$p_f = \alpha/(\alpha + \beta).$$

Let the average arrival rate and average service rate in steady state be denoted by $\hat{\lambda} = \lambda p_w + \lambda_1 p_f$ and $\hat{\mu} = \mu p_w$; furthermore, let $r = \hat{\lambda}/\hat{\mu}$. It will be demonstrated shortly that, as is often the case, $\hat{\lambda} < \hat{\mu}$ is a condition for steady state. Note that each of the quantities p_w, p_f, $\hat{\lambda}$, $\hat{\mu}$, and r would be independent of c even if the temporary assumption $c = 1$ were dropped.

The Generating Function

Generating-function techniques must be used to further analyze the model as there is no way of solving (1) in a recursive manner to obtain closed-form expressions for the $\{p_{ij}\}$. Define the generating functions

$$G_i(z) = \Sigma_{j=0}^{\infty} p_{ij}\, z^j \qquad |z| \leqslant 1,\ i = 0, 1, 2, \ldots, m$$

$$G(z) = \Sigma_{i=0}^{m} G_i(z) \qquad |z| \leqslant 1. \tag{2}$$

Multiplying each equation of the sets $\{(1a), (1d)\}$, $\{(1b), (1e)\}$, and $\{(1c), (1f)\}$ by z^j and summing over all j yields, respectively,

$$G_m(z) = [\alpha/(\lambda_1 + m\beta - \lambda_1 z^k)]\, G_0(z), \tag{3}$$

$$G_i(z) = [m\beta/(\lambda_1 + m\beta - \lambda_1 z^k)]\, G_{i+1}(z), \qquad (1 \leqslant i \leqslant m-1), \tag{4}$$

$$G_0(z) = [m\beta z\, G_1(z) - k\mu\, p_{00}(1 - z)]/[(\lambda + \alpha + k\mu)z - \lambda z^{k+1} - k\mu]. \tag{5}$$

Equations (3) and (4) can be used recursively to express $G_1(z)$ in terms of $G_0(z)$ as

$$G_1(z) = (\alpha/m\beta)\, [m\beta/(\lambda_1 + m\beta - \lambda_1 z^k)]^m\, G_0(z) \tag{6}$$

and (5) can be rearranged as

$$G_1(z) = (m\beta z)^{-1}\{[(\lambda + \alpha + k\mu)z - \lambda z^{k+1} - k\mu]\, G_0(z) + k\mu p_{00}(1 - z)\}. \tag{7}$$

Equating the expressions for $G_1(z)$ in (6) and (7) and solving for $G_0(z)$ gives

$$G_0(z) = \{k\mu\, p_{00}(1 - z)\, [f(z)]^m\}/D(z), \tag{8}$$

where

$$f(z) = 1 + (\lambda_1/m\beta)\, (1 - z^k), \tag{9}$$

and the denominator is

$$D(z) = \alpha z + [k\mu + \lambda z^{k+1} - (\lambda + \alpha + k\mu)z]\, [f(z)]^m. \tag{10}$$

Using (3) and (4) recursively results in

$$G_i(z) = \{(\alpha/m\beta)\ k\mu\ p_{00}(1-z)\ [f(z)]^{i-1}\}/D(z) \quad (1 \leqslant i \leqslant m). \tag{11}$$

$G(z)$ can now be calculated from (2), (8), and (11) as

$$G(z) = p_{00}[N(z)/D(z)], \tag{12}$$

where

$$N(z) = k\mu(1-z)\ \{[f(z)]^m + (\alpha/m\beta)\ \Sigma_{i=0}^{m-1}\ [f(z)]^i\}. \tag{13}$$

Using $G(1) = 1$ and computing $\lim_{z \to 1} G(z)$ by applying L'Hospital's rule to (12) yields, after algebraic simplification,

$$p_{00} = (1-r)\ p_w. \tag{14}$$

Substituting (14) into (12) results in a final expression for $G(z)$, that is

$$G(z) = (1-r)\ [\beta/(\alpha+\beta)]\ [N(z)/D(z)]. \tag{15}$$

Since $p_{00} > 0$, (14) also verifies the previously mentioned condition for steady state, $\hat{\lambda} < \ddot{\mu}$.

The busy fraction $\rho \equiv 1 - \Sigma_{i=0}^{m}\ p_{i0}$ equals $1 - G(0)$ and can be evaluated from (15), (9), (10), and (13). Provided $\lambda_1 > 0$,

$$\rho = 1 - (1-r)\ [\beta/(\alpha+\beta)]\ \{1 + (\alpha/\lambda_1)\ [1 - (m\beta/(\lambda_1 + m\beta))^m]\}.$$

In general, then, $\rho \neq (\hat{\lambda}/\hat{\mu})$; however, if $\lambda_1 = 0$, $\rho = (\hat{\lambda}/\hat{\mu})$ does hold.

Recursions for the $\{p_{ij}\}$

Unfortunately, no simple relationship exists relating the $\{p_{ij}\}$ to $p_{00}, p_{10}, \ldots, p_{m0}$. However, the $\{p_{ij}\}$ can be computed efficiently from (14) and the set of recursive equations (for $j = 0, 1, 2, \ldots$)

$$p_{mj} = (\lambda_1 + m\beta)^{-1}\ (\alpha p_{0j} + \lambda_1 p_{m,j-k}) \tag{16a}$$

$$p_{ij} = (\lambda_1 + m\beta)^{-1}\ (m\beta\, p_{i+1,j} + \lambda_1 p_{i,j-k}) \quad (i = m-1,\ m-2, \ldots, 1) \tag{16b}$$

$$p_{0,j+1} = (k\mu)^{-1}(\lambda\, \Sigma_{n=j+1-k}^{j}\ p_{0n} + \alpha\, \Sigma_{n=0}^{j}\ p_{0n} - m\beta\, \Sigma_{n=0}^{j}\, p_{1n}) \tag{16c}$$

where a term is zero if it has a negative subscript and the lower limit of summation in the first term of (16c) is reset to 0 if it is negative. Equations (16a) and (16b) are obviously equivalent to (1d) and (1e) while (16c) follows from (1f) and a simple inductive argument. Of course, the steady-state probability of having n customers in the system is given by $1 - \rho$ for $n = 0$ and, for $n > 0$, by $\Sigma_{i=0}^{m}\ \Sigma_{j=(n-1)k+1}^{nk}\, p_{ij}$.

Computation of L

The computation of L and L_q, the steady-state average number of customers in the system and in the queue, respectively, require some preliminary results. Let L^P and L_q^P be steady-state notation for the average number of customer service phases in the system and in the queue, respectively. Furthermore, let L_s^P denote the average number of service phases remaining for the customer (if any) in service. The following relationships clearly hold:

$$L_q^P = kL_q, \tag{17}$$

$$L^P = L_q^P + L_s^P \tag{18}$$

$$L = L_q + \rho. \tag{19}$$

Substituting (17) into (18) and solving for L_q results in

$$L_q = (1/k)\ (L^P - L_s^P). \tag{20}$$

Relationships (19) and (20) then yield

$$L = (1/k)\ (L^P - L_s^P) + \rho. \tag{21}$$

The problem, then, is to compute L^P and L_s^P.

Now $L^P = G'(1)$; however, evaluating $G'(1)$ is not easy. Expressing $G(z) = p_{00}$ $N(z)/D(z)$ and using L'Hospital's rule twice gives

$$G'(1) = p_{00}[N''(1) \cdot D'(1) - N'(1) \cdot D''(1)]/2[D'(1)]^2. \tag{22}$$

The algebraic manipulations required by (22) are straightforward but quite long. Because they would occupy several pages, the computations are omitted; however, they result in

$$L^P = \frac{r}{1-r}\left[\frac{k+1}{2} + \frac{m+1}{2m} \cdot \frac{k\alpha\lambda_1(\mu - \lambda + \lambda_1)}{\hat{\lambda}(\alpha+\beta)^2}\right]. \tag{23}$$

Obtaining L_s^P requires the development of another generating function. In the $M/E_k/1$ queue *not* subject to breakdown, given that a customer is in service in steady-state, the number of phases remaining until his service is complete is equally likely to be 1, 2, ... , or k (cf. [4, p. 169]). However, this is not the case when the queue is subject to breakdown. Define the generating functions

$$H_i(y) = p_{i0} + \Sigma_{n=1}^{k} (\Sigma_{j=0}^{\infty} p_{i,jk+n}) y^n, \qquad |y| \leqslant 1,\ i = 0, 1,\ \ldots,\ m$$

$$H(y) = \Sigma_{i=0}^{m} H_i(y), \qquad |y| \leqslant 1. \tag{24}$$

Clearly, $L_s^P = H'(1)$. As demonstrated in the Appendix, the lengthy derivation of $H(y)$ results in

$$H(y) = (1 - \rho) + (\rho - r)y^k + r\{[y(1 - y^k)]/[k(1 - y)]\}. \tag{25}$$

Application of L'Hospital's rule twice to (25) yields

$$L_s^P = H'(1) = (\rho - r)k + r[(k + 1)/2]. \tag{26}$$

Finally, combining (21), (23), and (26) results in

$$L = \frac{r}{1 - r}\left[(1 - r) + \frac{k + 1}{2k} \cdot r + \frac{m + 1}{2m} \cdot \frac{\alpha\lambda_1(\mu + \lambda_1 - \lambda)}{\hat{\lambda}(\alpha + \beta)^2}\right]. \tag{27}$$

Computation of *W*

The server may be idle for one or both of two reasons: the system is failed or no customers are present. Let b denote the steady-state probability that the server is *not* idle. Then the definitions of $\hat{\lambda}$, r, and p_w and expression (14) yield

$$b = p_w - p_{00} = \hat{\lambda}/\mu. \tag{28}$$

When the server is not idle, customers depart from the system at rate μ; of course, when the server is idle, no customers depart from the system. Hence, the average steady-state rate at which customers depart from the system is given by

$$\mu b = \hat{\lambda}. \tag{29}$$

Thus, in steady state, the average rate customers arrive at the system equals the average rate customers depart from the system.

Little's formula,

$$L = \hat{\lambda} W, \tag{30}$$

can now be used to compute W, the steady-state expected value of the time a customer spends both in the queue and in service. In particular,

$$W = L/\hat{\lambda} \tag{31}$$

where L is given by (27).

2. SPECIAL AND EXTREME CASES

CASE A. As $\beta \rightarrow \infty$, the repair periods have shorter and shorter durations, and, in the limit, repair is instantaneous. Intuitively, then, as $\beta \rightarrow \infty$, the model developed in Section 1 approaches the $M/E_k/1$ queue *not* subject to breakdown and having constant arrival rate λ and service rate μ. That this is in fact the case can be shown from (15), (27), and (31); as $\beta \rightarrow \infty$, $G(\cdot)$, L, and W all approach the corresponding quantities for the $M/E_k/1$ queue.

CASE B. If $k = 1$ and/or $m = 1$, exponential service and/or repair times result. When both $k = 1$ and $m = 1$, expression (27) for L reduces to a special case ($\mu_2 = 0$) of expression (33) of Yechiali and Naor [7, p. 729].

CASE C. Constant service and/or repair times can be analyzed by letting $k \rightarrow \infty$ and/or $m \rightarrow \infty$. It is clear from (27) and (31) that L and W are decreasing and convex functions of both k and m. Their limiting values are obtained by replacing $[(k+1)/2k]$ and/or $[(m+1)/2m]$ by 1/2 in (27) and (31).

CASE D. If the Poisson arrival rate is stationary ($\lambda_1 = \lambda$), expressions (27) and (31) for L and W reduce to special cases (Erlang service and repair) of relationships (24) and (26) of Avi-Itzhak and Naor [1, p. 309].

CASE E. In some practical situations, no customers enter the system when it is failed, either by their own choice or because of restrictions by the system. During the repair process, then, customers neither enter nor leave the system. Thus, it is intuitive that not only L but also the steady-state probabilities of having n customers in the system ($n = 0, 1, 2, \ldots$) are equal to those for the $M/E_k/1$ queue *not* subject to breakdown and having constant arrival rate λ and service rate μ. That this is in fact the case can be seen by setting $\lambda_1 = 0$. Then $r = \lambda/\mu$ and, from (15) and (27), both $G(z)$ and L simplify to the corresponding quantities for the $M/E_k/1$ queue. Of course, as (31) with $\lambda_1 = 0$ indicates, W is greater than in the $M/E_k/1$ case.

3. BEHAVIOR OF L AND W AS A FUNCTION OF c

In order to investigate how the system behaves as a function of c, the assumption of Sections 1, 2, and 4 that $c = 1$ is now dropped. Recall that c controls how rapidly the system oscillates between the operational and failed states. Varying c while holding α and β constant does not change p_w and p_f, the steady-state probabilities of the system being operational and failed, respectively. However, as c increases, the system fluctuates more rapidly between the operational and failed states, or, equivalently, the mean time the system stays in each state approaches 0.

Expressions for L and W as a function of c can be obtained by replacing α and β by $c\alpha$ and $c\beta$ everywhere in (27) and (31). From the expressions that result, it is easily shown that

$$L'(c) = -ac^{-2}$$

$$W'(c) = -bc^{-2}$$

where a and b are positive constants involving m, λ, λ_1, μ, α, and β. Hence, as functions of c, both L and W *decrease at a rate that is inversely proportional to* c^2.

To interpret this result qualitatively, consider two equally reliable systems (i.e., identical p_w's) also having identical k, m, λ, λ_1, and μ; however, suppose one system has infrequent failures but long repair times (a low c) and the other undergoes frequent but quickly repaired failures (a high c). If the objective is to minimize L and W, then the latter system should be chosen.

On a more quantitative level, the result supports a general conjecture of Ross [6] that, in a single server infinite capacity queueing model, the "more stationary" the Poisson arrival process is, then the smaller the average customer delay. As in [2] and [6], this conjecture has been verified in a special case. To see this, the behavior of the system as $c \rightarrow \infty$ will be investigated. It is easy to show from (27) and (31) that as $c \rightarrow \infty$, both L and W approach the corresponding quantities for an $M/E_k/1$ queue *not* subject to breakdown and having constant arrival rate $\hat{\lambda}$ and service rate $\hat{\mu}$. Thus, as $c \rightarrow \infty$, the system becomes more stationary in the sense that it behaves more and more like the $M/E_k/1$ queue with parameters $\hat{\lambda}$ and $\hat{\mu}$. Also, note that since L and W are decreasing in c, the smallest values they can ever achieve are the corresponding values for the $M/E_k/1$ queue with parameters $\hat{\lambda}$ and $\hat{\mu}$.

4. APPLICATION TO A PRODUCTION-STORAGE SYSTEM

By regarding the server as a production process turning out items one at a time and each customer as a unit demand for the product, the results of Section 1 can be used to analyze a production-storage process subject to breakdown and having the following additional characteristics:

(a) Unsatisfied demand is always backlogged.

(b) Items not needed immediately to satisfy backlogged demand are stored for future use up to a level of S, the finite capacity of the storage facility.

(c) When the storage facility is filled, no production takes place.

In such a production-storage model, three quantities of interest are I, the average number of items (physically) in inventory; B, the average number of backlogged items; and R, the fraction of time that demand can be met without backlogging. For example, it may be desired to choose S so that some or all of the conditions $I \geqslant c_1$, $B \leqslant c_2$, and $R \geqslant c_3$ hold, where c_1, c_2, and c_3 are specified design parameters. Meyer, Rothkopf, and Smith [5] consider four production-storage models differing from the one just described in that production and demand both occur at constant rates, unsatisfied demand is lost, and the durations of operating and failed periods have distributions corresponding to the four possible combinations of constant and exponential.

A state (i,j) of the queueing system translates into a state of the production-storage system as follows: $0 \leqslant j \leqslant kS$ is equivalent to $kS - j$ phases in storage and $j \geqslant kS$ is equivalent to $j - kS$ phases backlogged. Of course, i has the same interpretation in both models. Given this one-to-one correspondence between states in the two models, expression (27), recursions (16), and the observation that $B - I = L - S$ can be used to compute

$$I = S(1 - \rho) + \Sigma_{n=1}^{S-1} (S - n) (\Sigma_{i=0}^{m} \Sigma_{j=(n-1)k+1}^{nk} p_{ij})$$

$$B = L - S + I$$

$$R = \Sigma_{i=0}^{m} \Sigma_{j=0}^{k(S-1)} p_{ij} \quad (S > 0).$$

As a numerical example, consider a system with $\alpha = 1$, $\beta = 3$, $m = 5$, $\lambda = 8$, $\lambda_1 = 4$, $\mu = 16$, $k = 4$ and suppose it is desired to choose S so that $I \geqslant 3$, $B \leqslant 1$, and $R \geqslant 0.95$ all hold. Computed with the aid of (27) and (16), the following table shows I, B, and R as a function of S:

S	0	1	2	3	4	5	6	7	8	9	10
I	0.000	0.367	1.006	1.803	2.690	3.626	4.591	5.571	6.560	7.554	8.550
B	1.454	0.820	0.460	0.257	0.143	0.080	0.044	0.025	0.014	0.008	0.004
R	0.000	0.367	0.639	0.797	0.886	0.936	0.965	0.980	0.989	0.994	0.997

From the table, it is clear that $S \geqslant 6$ must hold.

ACKNOWLEDGMENTS

This research was supported by the Office of Naval Research under Contract N00014-75-C-0561 and was performed while the author was a visitor at the Department of Operations Research at Stanford University.

REFERENCES

[1] Avi-Itzhak, B. and P. Naor, "Some Queueing Problems with the Service Stations Subject to Breakdown," Operations Research, 11 303-320 (1963).

[2] Fond, S. and S. Ross, "A Heterogeneous Arrival and Service Queueing Loss Model," Technical Report ORC 77-12, Operations Research Center, University of California, Berkeley, CA, (May 1977).

[3] Gaver, D.P. Jr., "A Waiting Line with Interrupted Service, Including Priorities," Royal Statistical Society Journal, **B24**, 73-90 (1962).

[4] Gross, D. and C.M. Harris, *Fundamentals of Queueing Theory* (John Wiley and Sons, New York, 1974).

[5] Meyer, R.R., M.H. Rothkopf, and S.A. Smith, "Reliability and Inventory in a Production-Storage System," ARG Report 77-6, Xerox Palo Alto Research Center, Palo Alto, CA, January 1977 (forthcoming in Management Science).

[6] Ross, S., "Average Delay in Queues with Nonstationary Poisson Arrivals," Technical Report ORC 77-13, Operations Research Center, University of California, Berkeley, CA, (May 1977).

[7] Yechiali, U. and P. Naor, "Queueing Problems with Heterogeneous Arrivals and Service," Operations Research **19**, 722-734 (1971).

APPENDIX: DERIVATION OF $H(y)$

To obtain $H(y)$ multiply each equation of the sets {(1a), (1d)}, {(1b), (1e)}, and {(1c), (1f)} by y^n where

$$n = \begin{cases} 0 & \text{if } j = 0 \\ k & \text{if } j = k,\ 2k,\ 3k, \ldots \\ j \text{ modulo } k & \text{otherwise.} \end{cases}$$

Summing over all j then yields, respectively,

$$H_m(y) = (m\beta)^{-1}\,[\alpha H_0(y) - \lambda_1 p_{m0}(1 - y^k)] \tag{32}$$

$$H_i(y) = H_{i+1}(y) - (m\beta)^{-1}\lambda_1 p_{i0}(1 - y^k) \quad (1 \leqslant i \leqslant m - 1) \tag{33}$$

$$H_0(y) = [\alpha y - k\mu(1 - y)]^{-1}\,\{m\beta y\, H_1(y) - k\mu p_{00}(1 - y) - y(1 - y^k)$$

$$\cdot\,[\lambda p_{00} + k\mu(H_0^{'}(0) - p_{01})]\}. \tag{34}$$

Equations (32) and (33) can be used recursively to express $H_1(y)$ in terms of $H_0(y)$ as

$$H_1(y) = (m\beta)^{-1}\,[\alpha H_0(y) - \lambda_1(1 - y^k)\,\Sigma_{i=1}^{m} p_{i0}] \tag{35}$$

and (34) can be rearranged as

$$H_1(y) = (m\beta y)^{-1}\{[\alpha y - k\mu(1 - y)]\,H_0(y) + k\mu\, p_{00}(1 - y)$$

$$+\, y(1 - y^k)\,[\lambda p_{00} + k\mu(H_o^{'}(0) - p_{01})]\}. \tag{36}$$

Equating the expressions for $H_1(y)$ in (35) and (36) solving for $H_0(y)$ results in

$$H_0(y) = p_{00} + \{[y(1 - y^k)]/[k\mu(1 - y)]\}$$

$$\cdot\,\{\lambda_1\,\Sigma_{i=1}^{m} p_{i0} + \lambda p_{00} + k\mu[H_0^{'}(0) - p_{01}]\}. \tag{37}$$

Using (32) and (33) recursively yields

$$H_i(y) = (m\beta)^{-1}\,[\alpha H_0(y) - \lambda_1(1 - y^k)\,\Sigma_{n=i}^{m} p_{n0}], \ (1 \leqslant i \leqslant m) \tag{38}$$

$H(y)$ can now be obtained from (24), (37), and (38) as

$$H(y) = [(\alpha + \beta)/\beta]\left\{p_{00} + \frac{y(1-y^k)}{k\mu(1-y)}[\lambda p_{00} + \lambda_1 \Sigma_{i=1}^{m} p_{i0} + k\mu(H_0'(0) - p_{01})]\right\}$$

$$- (\lambda_1/m\beta)(1 - y^k)\Sigma_{i=1}^{m} ip_{i0}. \tag{39}$$

Using $H(1) = 1$ and computing $\lim_{y\to 1}H(y)$ by applying L'Hospital's rule to (39) results, after algebraic simplification, in

$$k\mu[H_0'(0) - p_{01}] = [(\beta/\alpha + \beta) - p_{00}]\mu - \lambda_1 \Sigma_{i=1}^{m} p_{i0} - \lambda p_{00}. \tag{40}$$

Substituing (40) into (39) gives

$$H(y) = [(\alpha + \beta)/\beta]p_{00} + r\{[y(1-y^k)]/[k(1-y)]\} - (\lambda_1/m\beta)(1-y^k)\Sigma_{i=1}^{m} ip_{i0} \tag{41}$$

Equations (1a) and (1b) can be rewritten as a

$$m\beta p_{m0} = \alpha p_{00} - \lambda_1 p_{m0}$$

$$m\beta p_{i0} = m\beta p_{i+1,0} - \lambda_1 p_{i0} \quad (1 \leqslant i \leqslant m - 1)$$

which is equivalent (by induction) to

$$p_{i0} = (m\beta)^{-1}(\alpha p_{00} - \lambda_1 \Sigma_{n=i}^{m} p_{n0}) \quad (i = m, m-1, \dots, 1). \tag{42}$$

Using (42) to evaluate $\Sigma_{i=0}^{m} p_{i0}$ yields the relationship

$$1 - \rho = [(\alpha + \beta)/\beta]p_{00} - (\lambda_1/m\beta)\Sigma_{i=1}^{m} ip_{i0}. \tag{43}$$

Because of (43) and (14), (41) simplifies to (25), the expression for $H(y)$ given in Section 1.

A COMPARISON OF WAITING TIME APPROXIMATIONS IN SERIES QUEUEING SYSTEMS

Daniel G. Shimshak

University of Massachusetts at Boston
Boston, Massachusetts

ABSTRACT

The determination of steady-state characteristics in systems of tandem queues has been left to computer simulation because of the lack of exact solutions in all but the simplest newtorks. In this paper, several methods developed for approximating the average waiting time in single-server queues are extended to systems of queues in series. Three methods, due to Fraker, Page, and Marchal, are compared along with results gathered through GPSS simulation. Various queueing networks with Erlangian service distributions are investigated.

INTRODUCTION

Series queueing systems, in which the departure process from one service station forms the arrival process at the next service station, are quite common in practice. Such systems have been used to represent inspection systems, production lines, telephone networks, registration processes, and urban traffic situations. However mathematical formulae only exist for problems that have restrictive assumptions associated with them. A relaxation of these assumptions results in problems that do not allow exact analytical solutions.

For a system with Poisson arrivals and exponential service times, R. R. P. Jackson [8] found the queue lengths of the service stations to be independent variables in the steady state. J. R. Jackson [7] demonstrated that, for this same Poisson-exponential system, the steady state joint probability distribution of customers waiting in the system is equal to the product of the probabilities for each individual Poisson-exponential service station.

Burke [1] showed that for each service station in the Poisson input-exponential service system, the steady state output process, and therefore the input process to the next station, is also Poisson. This proof was supplemented by Finch [4] who found Burke's Poisson departure to hold only when infinite queue lengths are allowed between stations. In addition he proved that successive interdeparture intervals are independent in the steady state only in the case of exponential service times and unbounded queue lengths. For general service distributions, other considerations are necessary to determine the departure process from each station in the series.

Because departures from one station form arrivals into the next station, the analysis of series queueing systems is much more complicated than traditional analysis of ordinary service stations. Most of the analytical work done with series queueing systems has been limited to Poisson-exponential networks. Further study has been performed through simulation. For example, Nelson [13] estimated steady-state queue statistics by simulating a two-server network model. He considered the exponential, Erlang with parameter 2, and constant distributions as arrival and service processess and conducted experiments for combinations of these.

Recent analysis in series queueing systems has turned to approximate solutions under steady state conditions. A useful contribution was made by Fraker [5] who experimentally developed an approximate formula for the mean waiting time in a system of single-server, infinite capacity queues with Erlang service. Page [14] developed an approximation for the average waiting time in $E_i/E_k/s$ queues and Marchal [10] did the same for $GI/G/$ queues. In this paper each of these will be extended to infinite capacity queues in series.

The purpose of this paper is to compare these three waiting time approximations with each other and with known analytic results, if they are available, or simulation results in the cases where analytic results do not exist. Four series queueing systems with diverse parameters are studied and diagramed in Figure 1. The intention is to show that approximation methods can be used effectively in the study and analysis of queueing systems in place of simulation whenever analytic solutions do not exist. The benefits in terms of savings in cost, time, and trouble are quite obvious.

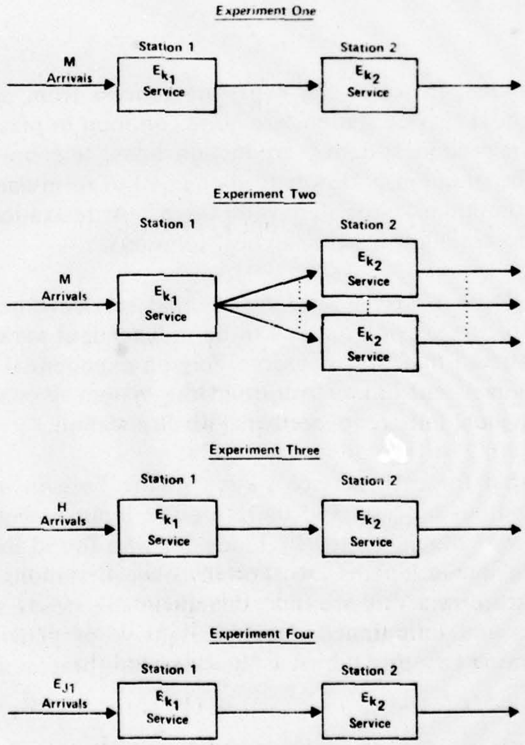


FIGURE 1. Diagrams of the four experimental queueing systems

The choice of the Erlang and hyperexponential families of probability distributions as arrival and service processes provides much greater modeling flexibility than does the exponential. In fact, the exponential is a special case of both the Erlang and hyperexponential. The Erlang distribution interpolates infinitely many distributions between the completely random exponential and completely regular constant. The coefficient of variation of the Erlang distributions ranges from 0 to 1. The hyperexponential represents distributions whose coefficient of variation ranges from 1 to ∞. Observing the mean and variance of the arrivals and services for some system would allow selecting a model to fit the system from either the Erlang or hyperexponential distributions. Thus the four systems chosen for study represent various combinations from a wide range of real-world systems.

MATHEMATICAL AND EXPERIMENTAL TECHNIQUES

The system considered has two service stations in series. Let λ_i, μ_i, and p_i be the arrival rates, service rates, and utilization rates for the two stations, for $i=1,2$. λ_1 is assumed to be known; under steady state conditions, the average departure rate from the first queue equals its average arrival rate, hence $\lambda_2 = \lambda_1 = \lambda$. The two stations have Erlang service distributions, independent of each other and of the arrival process, with parameters k_1 and k_2. The variances of the two service distributions are $\sigma_{s_i}^2 = 1/k_1\mu_i^2$ for $i=1,2$ respectively.

Marshall [11] derived a relationship for $GI/G/1$ queues which is often used in this paper. If σ_a^2, σ_s^2, and σ_d^2 are the variances of the interarrival, service, and departure distributions respectively, then the average waiting time in queue, $\overline{W}_q$, is

$$(1) \qquad \overline{W}_q = (\sigma_a^2 + 2\sigma_s^2 - \sigma_d^2)/2(1-\rho).$$

First we present the three approximation techniques by considering a single station system. Their application to service stations in series is explained in the next section as each experiment is discussed. In addition, the computer simulation techniques are described.

(i) *Fraker's Approximation*

For the $GI/E_k/1$ queueing system, σ_d^2 is not known exactly, but Fraker found the following approximate expression:

$$(2) \qquad \sigma_d^2 \simeq 1/j\lambda^2 + (j-1)j\mu^2 + (1-\rho)\,(j-1)/mj\mu^2 - (m-1)/m\mu^2$$
$$+\,0.5(1-\rho)\,(m-1)\,(j-1)/m^2 j\mu^2 + 2(1-\rho)\,(m-1)\,(j-1)/mj^2\mu^2$$

where j=(coefficient of variation of interarrival times)$^{-2}$ and m=(coefficient of variation of service times)$^{-2}$. Substituting σ_d^2 into (1), together with the known expressions for σ_a^2 and σ_s^2, yields the average waiting time in the queue.

(ii) *Page's Approximation*

Page developed an approximation for the average waiting time in the $E_j/E_k/s$ queue. Letting $\overline{W}_q\,(1/j,1/k)$ be the average waiting time in this queue, $\overline{W}_q\,(1,1)$=average waiting time in $M/M/s$, $\overline{W}_q(0,1)$ = average waiting time in D/M/s, $\overline{W}_q(1,0)$=average waiting time in M/D/s, and $\overline{W}_q(0,0)$=average waiting time in D/D/s, Page suggested that

$$\overline{W}_q(1/j,1/k) \simeq (1-1/j)\,(1-1/k)\,\overline{W}_q\,(0,0) + (1-1/j)\,(1/k)\,\overline{W}_q(0,1)$$
$$+\,(1/j)\,(1-1/k)\,\overline{W}_q\,(1,0) + (1/j)\,(1/k)\,\overline{W}_q\,(1,1).$$

Since $\overline{W}_q(0,0)=0$, this becomes

$$(3) \qquad \overline{W}_q \simeq (1-1/j)\,(1/k)\,\overline{W}_q\,(0,1) + (1/j)\,(1-1/k)\,\overline{W}_q\,(1,0) + (1/j)\,(1/k)\,\overline{W}_q(1,1).$$

(iii) *Marchal's Approximation*

Marchal found an approximate formula for waiting time in $GI/G/1$ queues to be

$$\overline{W}_q \simeq [(1+1/m)/(1/\rho^2+1/m)]\ [\lambda(\sigma_a^2+\sigma_s^2)/2(1-\rho)] \tag{4}$$

where $1/m$, as before, is the square of the coefficient of variation of the service times.

(iv) *Simulation Techniques*

Simulation results were obtained through GPSS simulation of the queueing system using the regenerative method suggested by Crane and Iglehart [2]. Independent and identically distributed blocks of observations are created by starting the simulation in the empty state and letting it return to the empty state. All observations, transient and steady state ones, are averaged within these regeneration cycles. Confidence intervals are found using standard statistical formulae. For the systems considered in this paper, the number of customers simulated ranged from 15,000 to 25,000 depending upon the utilization rate of the service stations.

It should be pointed out that the approximation techniques used here all assume that the input to the second station in the series is a renewal process. This is not true. Results of Finch [4] indicate that departure intervals are statistically dependent random variables for any system other than those with Poisson arrivals and exponential service times. Later Disney, Farrell, and De Morais [3] determined that the departure process from an $M/G/1$ queue is a renewal process for some additional simple systems. For the series queueing systems considered in this paper, arrival intervals at the second station are statistically dependent. However, the methods applied here are only approximations, and they do not account for the lack of independence in the input and output processes.

A report by Hillier and Lo [6] includes tables of statistics for various $E_j/E_k/s$ systems with small integer values of j,k, and s. These numerical results, obtained through extensive computational work, do not include any of the systems observed in this paper. In addition they suggest an approximation for other systems, which is somewhat restrictive, by extrapolating out from the existing tables of data. This approximation is not applicable here since the parameters for the Erlang arrivals into the second station are not any of the integer values considered in the tables. The results of Hillier and Lo can be useful in further analysis of approximation techniques.

RESULTS

Four series queueing systems were studied. The specific experimental design was decided upon with the idea of choosing parameters representing systems that were realistic while at the same time selecting diverse sets of parameters. In addition, the computing time served as a constraint on the number of experiments that were feasible. For all simulation, the mean interarrival time, $1/\lambda$, was held constant at 1 time unit of the simulator. This made determination and control of utilization rates an easy task.

Experiment One — $M/E_{k_1}/1 \rightarrow \cdot/E_{k_2}/1$

The queueing system at the first station is a well known case of the $M/G/1$ system. In a general $M/G/1$ system, the average waiting time is known from the Pollaczek-Khintchine formula to be $\rho(\mu^2\sigma_s^2 + 1)/2(1 - \rho)\mu$. If G is an Erlang with parameter k_1, then

$$\overline{W}_{q_1} = \rho_1(k_1+1)/2(1-\rho_1)k_1\mu_1. \tag{5}$$

Each of the three approximations yields this exact result at the first station. Furthermore, for the first station, the variance of the interdeparture times is known. Jenkins [9] proved it to be

$$\sigma_{d_1}^2 = [1-(k_1-1)\rho_1^2/k_1]/\lambda_1^2. \tag{6}$$

Since the output from stage 1 is the input for stage 2, $\sigma_{a_2}^2 = \sigma_{d_1}^2$. However the complete distribution of the interdeparture times is not known. Thus the second station is a $G/E_{k_2}/1$ system. G is the interarrival distribution, equal to the departure distribution from the first station, with mean $1/\lambda$ and variance $\sigma_{a_2}^2$ given in (6). Exact formulae for the average waiting time for this type of system are not available, thus $\overline{W}_{q2}$ must be computed through approximations. The total waiting time in the system is $\overline{W}_{q_T} = \overline{W}_{q_1} + \overline{W}_{q_2}$.

Using Fraker's technique, $\sigma_{d_2}^2$ is approximated by (2) using $m_2 = k_2$ and $j_2 = [1-(k_1-1)\ \rho_1^2/k_1]^{-1}$ which follows from (6). Substituting $\sigma_{d_2}^2$ in (1), where $\sigma_{a_2}^2$ and $\sigma_{s_2}^2$ are known, gives the following expression:

$$\begin{aligned}\overline{W}_{q2} \simeq {} & [\lambda/2(1-\rho_2)]\ [2/k_2\mu_2^2 - (j_2-1)/j_2\mu_2^2 - (1-\rho_2)\ (j_2-1)/k_2\, j_2\, \mu_2^2 \\ & + (k_2-1)\ k_2\, \mu_2^2 - 0.5(1-\rho_2)\ (k_2-1)\ (j_2-1)/k_2^2 j_2\, \mu_2^2 \\ & -2(1-\rho_2)\ (k_2-1)\ (j_2-1)/k_2 j_2^2 \mu_2^2].\end{aligned} \tag{7}$$

In applying Page's approximation, it must be noted that the second station in the series system is not strictly an $E_j/E_k/1$ queue, but since (3) is an approximation, it is likely that it may be satisfactory for the $G/G/1$ system at the second station. In order to apply (3) to the calculation of $\overline{W}_{q2}$, we set $j = [1-(k_1-1)\ \rho_1^2/\ k_1]^{-1}$ as before, and $k=k_2$. The waiting times of the simple systems with single servers are $\overline{W}_q(1,1) = \rho_2/(1-\rho_2)\mu_2$, $\overline{W}_q(0,1) = \nu/(1-\nu)\mu_2$ where ν is the root of $\exp[-(1-\nu)/\rho]$ in (0,1) and $\overline{W}_q(1,0) = \rho_2/2(1-\rho_2)\mu_2$. The Marchal approximation is applied to the second station with $m=k_2$ and the known variances substituted in (4).

Table I compares the results calculated by the three approximation techniques with those obtained through simulation, where k_1=10 and k_2 is allowed to vary. For this experiment alone, the waiting time at each of the stations along with the total average waiting time is reported. In this way the relative proportion of the total delay encountered at each station can be seen.

Experiment Two — $M/E_{k_1}/1 \rightarrow \cdot/E_{k_2}/s$

As in Experiment One, the system at the first station is well known and the waiting time is given in (5). A procedure is suggested by Rosenshine and Chandra [15] for approximating the waiting time at the second queue using Fraker's formula. They claim that (7) can be viewed as the average waiting time of the $M/M/1$ system multiplied by a factor Y, where

$$\begin{aligned}Y = \frac{1}{2}[& 2/k_2-(j_2-1)/j_2-(1-\rho_2)\ (j_2-1)/k_2 j_2 + (k_2-1)/k_2 \\ & - 0.5(1-\rho_2)\ (k_2-1)\ (j_2-1)/k_2^2 j_2 - 2\ (1-\rho_2)\ (k_2-1)\ (j_2-1)/k_2 j_2^2].\end{aligned} \tag{8}$$

Assume that (8) gives the ratio of the average waiting time at station 2 with s servers and general arrivals to the average waiting time with s servers and Poisson arrivals. To find the average waiting time at station 2 with multiple servers, calculate Y using (8) and multiply it by the average waiting time of an $M/M/s$ system with the same utilization rate.

The application of Page's method is similar to that in Experiment One where the second station is now approximated by an $E_j/E_k/s$ queue using (3). Marchal's approximation is not applicable in this case since it is defined only for single channel queues. Table II compares the approximate and simulation results, where k_1 = 10, k_2 varies, and s=10.

TABLE I — *Average Waiting Time for $M/E_{10}/1 \rightarrow \cdot/E_{k_2}/1$ System*

P_1	P_2	k_2	$\overline{W}_{q1}$†	Fraker		Page		Marchal		Simulation	
				$\overline{W}_{q2}$	$\overline{W}_{qT}$	$\overline{W}_{q2}$	$\overline{W}_{qT}$	$\overline{W}_{q2}$	$\overline{W}_{qT}$	Point Estimate $\overline{W}_{qT}$	95% Confidence Interval $\overline{W}_{qT}$
0.60	0.60	1	0.495	0.696	1.191	0.702	1.197	0.686	1.181	1.202	[1.111,1.290]
0.80	0.60	1	1.760	0.537	2.297	0.547	2.307	0.519	2.279	2.265	[2.031,2.476]
0.60	0.60	10	0.495	0.270	0.765	0.344	0.839*	0.340	0.835	0.779	[0.720,0.838]
0.80	0.60	10	1.760	0.142	1.902	0.226	1.986	0.220	1.980	1.828	[1.604,2.031]
0.60	0.80	1	0.495	2.578	3.073	2.602	3.097	2.568	3.063	3.414	[2.985,3.840]
0.80	0.80	1	1.760	2.094	3.854	2.137	3.897	2.076	3.836	4.325	[3.722,4.850]
0.60	0.80	10	0.495	1.100	1.595*	1.234	1.729	1.224	1.719	1.925	[1.657,2.192]
0.80	0.80	10	1.760	0.671	2.431	0.824	2.584	0.807	2.567	2.478	[2.185,2.740]

†All three approximations yield this exact result.
*Indicates that the approximation lies outside the confidence interval.

TABLE II — *Total Average Waiting Time* ($\overline{W}_{qT}$) *for* $M/E_{10}/1 \rightarrow \cdot/E_{k_2}/10$ *System*

ρ_1	ρ_2	k_2	Fraker	Page	Simulation	
					Point Estimate	95% Confidence Interval
0.60	0.60	1	0.507	0.506	0.518	[0.492,0.543]
0.80	0.60	1	1.769	1.767	1.764	[1.550,1.958]
0.60	0.60	10	0.500	0.502	0.513	[0.469,0.557]
0.80	0.60	10	1.762	1.764	1.767	[1.545,1.968]
0.60	0.80	1	0.627	0.621	0.655	[0.618,0.690]
0.80	0.80	1	1.867	1.857	1.899	[1.684,2.094]
0.60	0.80	10	0.551*	0.561*	0.598	[0.568,0.628]
0.80	0.80	10	1.794	1.803	1.861	[1.580,2.001]

*Indicates that the approximation lies outside the confidence interval.

Experiment Three — $H/E_{k_1}/1 \rightarrow \cdot/E_{k_2}/1$

The hyperexponential distribution, as defined by Morse [12], has coefficient of variation, c, greater than 1. Only Fraker and Marchal's approximations can be applied to this system.

Using Fraker's approximation, $\sigma^2_{d_1}$ is found from (2) where $j_1 = 1/c^2$ and $m_1 = k_1$. Knowing that $\sigma^2_{a_1} = c^2/\lambda^2$ and $\sigma^2_{s_1} = 1/k_1\mu_1^2$, $\overline{W}_{q_1}$ is found using (1). Since $\sigma^2_{a_2}=\sigma^2_{d_1}$, j_2 can be calculated. For $m_2 = k_2$, $\sigma^2_{d_2}$ is found, and with known expressions for $\sigma^2_{a_2}$ and $\sigma^2_{s_2}$, $\overline{W}_{q2}$ is also determined using (1).

Marchal's approximation yields $\overline{W}_{q1}$ when the appropriate expressions are substituted into (4). With the waiting time approximated, and $\sigma^2_{a_1}$ and $\sigma^2_{s_1}$ known, (1) can be used to solve for $\sigma^2_{d_1}$. This, in turn, defines $\sigma^2_{a_2}$, and with other values known, (4) determines $\overline{W}_{q_2}$.

In this experiment, the arrivals into the first station are hyperexponential with $c=2$. These results are presented in Table III. For the systems where the first station had Erlang service with parameter 10, Fraker's approximation gave negative waiting times at the second station. This failure in the approximation formula is indicated in the table with blanks.

TABLE III — *Total Average Waiting Time* ($\overline{W}_{qT}$) *for* $H/E_{k_1}/1 \rightarrow \cdot/E_{k_2}/1$ *System*

p_1	k_1	p_2	k_2	Fraker	Marchal	Simulation	
						Point Estimate	95% Confidence Interval
0.80	1	0.60	1	10.298*	10.389*	9.077	[7.688,10.180]
0.80	10	0.60	1	-----	7.913*	6.487	[5.762,7.083]
0.80	1	0.60	10	10.169*	9.863*	8.545	[7.354,9.541]
0.80	10	0.60	10	-----	7.427*	6.008	[5.276,6.613]
0.80	1	0.80	1	13.496	13.539	12.313	[10.053,13.691]
0.80	10	0.80	1	-----	10.779*	9.642	[8.313,10.695]
0.80	1	0.80	10	12.546*	11.903	11.127	[9.747,12.249]
0.80	10	0.80	10	-----	9.209*	7.401	[6.451,8.177]

*Indicates that the approximation lies outside the confidence interval. Blanks indicate a failure in the approximation formula to yield values.

Experiment Four — $E_{j_1}/E_{k_1}/1 \rightarrow \cdot/E_{k_2}/1$

Fraker's formula (2) is used for a given value of j_1 and $m_1 = k_1$ to determine $\sigma^2_{d_1}$. Here $\sigma^2_{a_1} = 1/j_1\lambda^2$ and $\sigma_{s_1} = 1/k_1\mu_1^2$ so that $\overline{W}_{q1}$ is found from (1). This provides the appropriate information to find $\overline{W}_{q2}$ as in Experiment Three.

Page's approximation in (3) yields $\overline{W}_{q1}$. As before, with $\overline{W}_{q1}$ approximated, σ^2_{a1} and σ^2_{s1} known, (1) is used to determine $\sigma^2_{d_1}$. Page approximates the arrivals to the second station by an E_j distribution with variance $1/j_2\lambda^2$. Since $\sigma^2_{d_1} = \sigma^2_{a2}$, j_2 can be found and used to calculate $\overline{W}_{q2}$. For Marchal's approximation, (4) is applied to find $\overline{W}_{q1}$. This is used together with (1) to yield $\sigma^2_{d_1}$ as in Experiment Three, which then allows determination of $\overline{W}_{q_2}$ using (4) again. Table IV shows the approximations together with simulated results. Here the parameter of the Erlangian arrival distribution into the first station, j_1, is 10.

TABLE IV — *Total Average Waiting Time* ($\overline{W}_{qT}$) *for* $E_{10}/E_{k_1}/1 \rightarrow \cdot/E_{k_2}/1$ *System*

p_1	k_1	p_2	k_2	Fraker	Page	Marchal	Simulation	
							Point Estimate	95% Confidence Interval
0.80	1	0.60	1	2.240	2.295	2.213	2.303	[2.115,2.471]
0.80	10	0.60	1	0.583	0.652*	0.589	0.588	[0.547,0.626]
0.80	1	0.60	10	1.814	1.924	1.845	1.953	[1.500,2.366]
0.80	10	0.60	10	0.272*	0.377*	0.346*	0.252	[0.236,0.267]
0.80	1	0.80	1	4.271*	4.304*	4.258*	3.840	[3.512,4.116]
0.80	10	0.80	1	1.770	1.853	1.753	1.822	[1.631,1.995]
0.80	1	0.80	10	2.793	2.916	2.877	2.684	[2.160,3.155]
0.80	10	0.80	10	0.498*	0.610*	0.575*	0.456	[0.433,0.479]

*Indicates that the approximation lies outside the confidence interval.

CONCLUSIONS AND RECOMMENDATIONS

Comparative experiments were run using three approximation formulae for the waiting time in the two-stage series queueing system. Table V shows the relative performance of each of the approximation techniques by comparing it to the waiting time found through computer simulation. For each experiment, the % deviations and average absolute % deviations are shown. Table VI summarizes each of the techniques in terms of overall average absolute % deviation and the percentage of times the approximate value fell within the 95% confidence

TABLE V — *% Deviation in Waiting Time of Approximate Value from Simulated Value*

Run	EXPERIMENT 1			EXPERIMENT 2		
	Fraker	Page	Marchal	Run	Fraker	Page
1	−0.96	−0.48	−1.82	1	−2.16	−2.36
2	1.40	1.85	0.59	2	0.30	0.19
3	−1.79	7.73	7.25	3	−2.58	−2.16
4	4.05	8.69	8.32	4	−0.25	−0.14
5	−9.98	−9.28	−10.28	5	−4.31	−5.16
6	−10.90	−9.91	−11.31	6	−1.67	−2.19
7	−17.12	−10.20	−10.70	7	−7.83	−6.19
8	−1.90	4.28	3.59	8	−3.57	−3.09
Average Absolute % Deviation	6.01	6.55	6.73		2.83	2.69

Run	EXPERIMENT 3		EXPERIMENT 4			
	Fraker	Marchal	Run	Fraker	Page	Marchal
1	13.46	14.46	1	−2.70	−0.34	−3.88
2	-----	21.99	2	−0.79	11.02	0.16
3	19.00	15.42	3	−7.09	−1.46	−5.53
4	-----	23.62	4	8.21	49.97	37.32
5	9.61	9.95	5	11.23	12.09	10.90
6	-----	11.79	6	2.82	1.72	3.76
7	12.75	6.97	7	4.05	8.64	7.18
8	-----	24.42	8	9.11	33.68	25.94
Average Absolute % Deviation	13.71	16.08		5.75	14.87	11.83

Blanks indicate a failure in the approximation formula to yield values.

TABLE VI — *Summary of Approximation Techniques for Experiments 1 and 4*

Technique	Overall Average Absolute % Deviation	% in Confidence Interval
Fraker	5.88	75.00
Page	10.71	68.75
Marchal	9.28	81.25

interval determined by simulation. Only the results of Experiments 1 and 4 are included in this table, since they cover all three techniques.

Each of the techniques was shown to have its limitations. Page's formula could not be applied to hyperexponential arrivals, Marchal's limited to a single channel at any stage, and Fraker's broke down under certain cases where the system had hyperexponential arrivals. However, it is apparent from the results shown here that the approximations can be successfully used in studying queueing systems when analytic solutions are not available. Overall, one can conclude that for series queueing systems with either exponential or Erlang interarrivals and service times, Fraker's approximation formula yields the best results. This is true for both single and multiple servers. When the queueing system has hyperexponential arrivals, Marchal's formula should be applied to obtain approximate results.

If Experiment Three is disregarded, the values obtained from each of the approximations are very close to each other, and, in most cases, consistent with respect to their position in the confidence interval found through simulation. It appears as if the three approximations either underestimate or overestimate the actual value. This could be due to the influence of statistical dependency in the service stations' arrival and departure processes. Investigation of the waiting time data gathered in each of the four experiments failed to yield any conclusive statements about this dependency. It would be interesting to see if these consistencies in the approximation results are general to other cases as well. Any future work that can account for the presence of statistical dependence in the input and output of the service stations will certainly lead to improved approximation techniques.

All too often the systems analyst resorts to simulation to study queueing systems when analytic solutions do not exist. Approximation techniques have been shown to be a useful tool in the study of queueing networks and can be used for deriving information on all system performance measures. Future research should develop better approximations that have application to all types of complex queueing systems and ultimately provide an attractive alternative to simulation. Approximation techniques should be put to increasing use, especially in the application of queueing analysis to real world situations involving complex queueing systems.

REFERENCES

[1] Burke, P. J., "The Output of a Queueing System," Operations Research 4, 699-704 (December 1956).

[2] Crane, M. A. and D. L. Iglehart, "Simulating Stable Stochastic Systems, I: General Multiserver Queues," Journal of the Association of Computing Machinery 21, 103-113 (January 1974).

[3] Disney, R. L., R. L. Farrell, and P. R., De Morais, "A Characterization of M/G/1 Queues with Renewal Departure Processes," Management Science 19, 1222-1228 (July 1973).

[4] Finch, P. D., "The Output Process of the Queueing System M/G/1," Journal of the Royal Statistical Society, Ser. B 21, 375-380 (1959).

[5] Fraker, J. R., "Approximate Techniques for the Analysis of Tandem Queueing Systems," Ph.D. dissertation, Clemson Univ. (1971).

[6] Hillier, F. S. and F. D. Lo, "Tables for Multiple-Server Queueing Systems Involving Erlang Distributions," Technical Report No. 31, Department of Operations Research, Stanford Univ. (December 1971).

[7] Jackson, J. R., "Networks of Waiting Lines," Operations Research 5, 518-521 (August 1957).

[8] Jackson, R. R. P., "Queueing Systems with Phase-Type Service," Operational Research Quarterly 5, 109-120 (December 1954).

[9] Jenkins, J. H., "On the Correlation Structure of the Departing Process of the M/E$_\lambda$/1 Queue," Journal of the Royal Statistical Society, Ser. B 28, 336-344 (1966).

[10] Marchal, W. G., "An Approximate Formula for Waiting Time in Single Server Queues," AIIE Transactions 8, 473-474 (December 1976).

[11] Marshall, K. T., "Some Inequalities in Queueing," Operations Research 16, 651-665 (May-June 1968).

[12] Morse, P. M., *Queues, Inventories and Maintenance* (John Wiley & Sons, New York, 1958).

[13] Nelson, R. T., "A Simulation Study and Analysis of a Two Station, Waiting-Line Network Model," Ph.D. dissertation, UCLA (1965).

[14] Page, E., *Queueing Theory in OR* (Crane Russak & Co., New York, 1972).

[15] Rosenshine, M. and M. J. Chandra, "Approximate Solutions for Some Two-Stage Tandem Queues, Part 1: Individual Arrivals at the Second Stage," Operations Research 23, 1155-1166 (November-December 1975).

STATISTICAL TESTS FOR EXPONENTIAL SERVICE FROM M/G/1 WAITING-TIME DATA*

T.R. Thiagarajan**

Washington Gas Light Co.
Washington, D.C.

Carl M. Harris

Mathematica, Inc.
Washington, D.C.

ABSTRACT

From an original motivation in quantitative inventory modeling, we develop methods for testing the hypothesis that the service times of an M/G/1 queue are exponentially distributed, given a sequence of observations of customer line and/or system waits. The approaches are mostly extensions of the well-known exponential goodness-of-fit test popularized by Gnedenko, which results from the observation that the sum of a random exponential sample is Erlang distributed and thus that the quotient of two independent exponential sample means is F distributed.

INTRODUCTION

As is discussed, for example, in the final chapter of Hadley and Whitin [8], it is imperative that data be collected for any quantitative inventory analysis in order to obtain (possibly estimated) values of model parameters and functions before it is at all feasible to find an "optimal" operating doctrine. Even if there is a satisfactory mathematical model available, its solution may be prevented by an inability to observe, much less calculate, its key variables.

It turns out that numerous inventory systems can be formulated as functions of a reorder queueing problem. That is, the reorder fulfillment facility acts as a server to a demand stream related or even identical to the underlying demand on the inventory, as shown in Figure 1. If, for example, the inventory manager is able only to observe the amounts of time it takes for reorders to return to stock (i.e., the leadtimes), then it is likely that a solution to the inventory model will not be possible. By virtue of the presentation of the queueing subsystem, this would then be analogous to trying to solve a waiting-line model given information only on its sequence of customer waiting times.

In some earlier work by Gross, Harris, and Lechner [7], and Gross and Harris [5], [6], stochastic inventory models were studied which made use of the relationship between the

*This work was supported by the Office of Naval Research under Contract No. N00014-67-A-0214, while the authors were with The George Washington University. Special thanks to the referees for many helpful suggestions.

**This author's contribution was based upon a dissertation submitted to the School of Engineering and Applied Science of the George Washington University in partial fulfillment of the requirements for the degree of Doctor of Science.

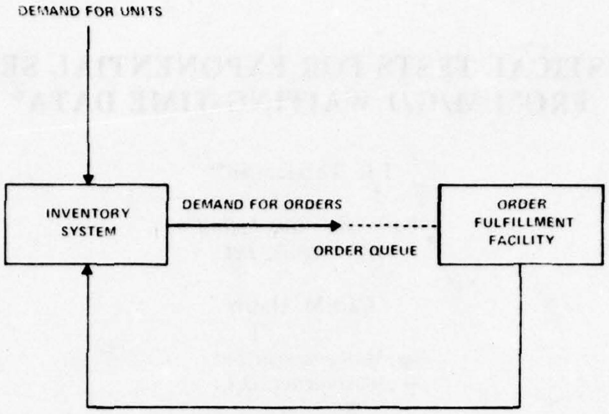


FIGURE 1. Inventory system

reorder queue and the overall inventory system. For example, the authors of [5] study several one-for-one ordering (continuous-review (s,S) policies where $s = S-1$) inventory models in which the time required for order replenishment, or the reorder leadtime, depends on the number of orders outstanding. Demand on the inventory system is assumed to be a Poisson process with constant mean rate, and the model is then solved.

When the continuous review $(S-1, S)$ problem is examined more carefully, we see a direct one-to-one relationship between the inventory state probabilities and those of the reorder queue, and it is easy to see that the reorder leadtimes correspond to the system waiting times of an M/G/1 queue. In order to complete the inventory analysis, it thus becomes essential to determine the nature of the arbitrary distribution associated with order refills, that is, to find G, which may have to be done using only successive values of the leadtime or some other similar empirical data. Therefore it seems that a very natural problem which may arise in this context is the determination of a test (or tests) of the hypothesis that the order fulfillment facility functions according to specified distribution such as the exponential. But this is precisely the problem of testing the hypothesis that $G = M$ for an M/G/1 queue given a sequence of observed waiting times. Of course, the analysis detailed in the following could be used to evaluate the service mechanism of any M/G/1 queue, whether from an inventory model or otherwise.

We do not mean to imply by our model and its analysis that we are now able to handle statistically the most complex queueing and inventory structures. Rather, this is but one step to provide the capability of making statistical inferences about probabilistic models in operations research given very limited information on their system behavior.

THE ANALYSIS

The object of this paper then is the development of methods for testing the hypothesis that the service times of an M/G/1 queue are exponentially distributed, given a sequence of observations of customer line and/or system waits. It turns out that the approaches are mainly variations on the well-known exponential test theme popularized by Gnedenko (for example, see [4]).

The Gnedenko test is as follows. Suppose $(t_1, t_2, \ldots, t_n)$ is a random sample of size n from a hypothesized exponential distribution, and let $(x_1, x_2, \ldots, x_n)$ be the order statistics obtained from the original sample. Then we define the i th normalized spacing as

$$s_i = (n-i+1)\,(x_i - x_{i-1}) \qquad \text{with } x_0 = 0.$$

It is a well-known result that the s_i are exponentially distributed (with the same parameter) whenever the t_i are. Therefore, after the data are split into two groups of sizes (r) and ($n-r$),

$$Q(r,n-r) = \frac{\sum_{i=1}^{r} s_i/r}{\sum_{i=r+1}^{n} s_i/(n-r)} \tag{1}$$

is distributed as an F with $2r$ and $2(n-r)$ degrees of freedom.

The proof of this assertion is fairly simple. We have the ratio of two Erlang random variables in (1). But an Erlang variable can be reduced to a chi-square variable by a suitable transformation and the ratio of two chi-square random variables divided by their degrees of freedom follows an F distribution, in this case with $2r$ and $2(n-r)$ degrees of freedom. We may therefore test for exponentiality by requiring $Q(r,n-r)$ to fall within an appropriate α-level rejection region based on the F. Furthermore, this approach can still be used even in the event the observed data come from a censored or hypercensored sample, as explained in [1], [4], and [9]. This test has been shown to be quite powerful against an extremely wide range of feasible alternative hypotheses ([3], [9]), at worst requiring a careful choice of r for splitting the data. Special guidance for choosing r is provided in [9] and [11] where the emphasis is on using the shape of the alternative's hazard rate.

For purposes of analysis, this work has been broken up into two main problems and then further into a number of subproblems. The organization is thus as follows:

I. Line wait data available

A. No line waits observed to be zero

1. no parameters known
2. at least one parameter known

B. Some line waits zero

1. no parameters known
2. at least one parameter known

II. System wait data available

A. No line waits zero

B. Some line waits zero

C. No information whatsoever regarding line waits

For all problems, the queues are assumed to be stationary but possibly in heavy traffic and the input stream Poisson with unknown rate λ unless specified otherwise.

The Line-Wait Problem (I)

To begin, suppose we have a queue and are able to observe the line waits of successive customers, say $W_q^{(1)}$, $W_q^{(2)}$, $\ldots$, $W_q^{(n)}$, and also know that the interarrival times are independent and identically distributed exponential random variables. It is then desired to test the hypothesis that the service distribution is exponential. In short, given the line waits for an M/G/1 queue, how do we test the hypothesis that $G = M$, in the notation of queueing? The main difficulty that arises is the fact that the line waits of any queue are correlated to each other, so we must find a way to remove the correlation first before we can deal with these types of data in any reasonably conventional way. The problem is even more complicated if some of the line waits are zero, which is naturally to be expected.

For any single channel FIFO (first-in, first-out) queue, the relationship between the line waiting times of the n th and $(n+1)st$ customers, say $W_q^{(n)}$ and $W_q^{(n+1)}$, is given by

$$W_q^{(n+1)} = \max\,(0, W_q^{(n)} + S^{(n)} - T^{(n)}) \tag{2}$$

where $S^{(n)}$ is the service time of the n th customer and $T^{(n)}$ is the time between the arrivals of the n th and $(n+1)st$ customers. We can note immediately that the stochastic process $\{W_q^{(n)}, n = 1, 2, 3, \ldots, \}$ is a discrete-time Markov process since the behavior of $W_q^{(n+1)}$ is only a function of the stochastically determined value of $W_q^{(n)}$, and is independent of the prior waiting-time history. Thus the successive first differences of the process will be independent and identically distributed (IID).

CASE A: Let us first discuss the simple case in which none of the line waits $W_q^{(n)}$, $n=1, 2, 3, \ldots$, is zero (and cannot be, as in extremely heavy traffic). Then from Equation (2) we have

$$W_q^{(n+1)} - W_q^{(n)} = S^{(n)} - T^{(n)}, \tag{3}$$

where the quantities $\{S^{(n)} - T^{(n)}\}$ are thus IID. Under the null hypothesis of $G = M$, the queue is an M/M/1, and since the interarrival and service times are exponentially distributed, the densities of T and S would be given by

$$\begin{bmatrix} f_1(t) = \lambda e^{-\lambda t}\,(t > 0) \\ f_2(s) = \mu e^{-\mu s}(s > 0) \end{bmatrix}. \tag{4}$$

Then the marginal distribution function of $Y=S-T$ can be derived as

$$G(y) = \begin{cases} \dfrac{\mu}{\lambda+\mu} e^{\lambda y} & (y \leqslant 0) \\ 1 - \dfrac{\lambda}{\lambda+\mu} e^{-\mu y} (y > 0) \end{cases} \tag{5}$$

Hence it follows that if we are given the line waits of successive customers from the M/G/1 queue, we can test the hypothesis $G = M$ by taking the first differences of the line waits and testing whether these differences have come from the distribution $G(y)$ given by Equation (5).

It is easy to see from Equation (5) that[1]

$$Pr\,\{Y > 0\} = \lambda/(\lambda+\mu) \tag{6}$$

and

$$Pr\,\{Y < 0\} = \mu/(\lambda+\mu) \tag{7}$$

and thus that the conditional densities of Y, given that $y > 0$ or $y < 0$, are

(8)
(9)

$$\begin{cases} g(y\,|\,y>0) = \mu\ \exp[-y\mu] \\ g(y\,|\,y<0) = \lambda\ \exp[y\lambda] \end{cases}$$

Equations (8) and (9) suggest that if we are given a sample and asked to test whether the sample could have come from the distribution of the difference of two exponentials, we can split the data into two groups, one consisting of the positive observations and the other consisting of the negative numbers;[2] then we test for exponentiality separately, using the F test for each, setting the Type I error with appropriate care.

To be somewhat more specific on the formalities of the testing process, it should be noted that though there are two separate independent exponential tests, there is really only one true hypothesis to'be tested, namely, that *both* exponentials are valid, with the alternative that either one or both of the sub-hypotheses is invalid. If α is the Type I error, then we see that

$$\begin{aligned} \alpha &\equiv Pr\,\{\text{rejection of } H_0\,|\,H_0 \text{ true}\} \\ &= Pr\{\text{rejection of either or both of the exponentials}|\text{both true}\} \\ &= Pr\{\text{rejection of first (say } g_1)\,|\, g_1 \ \&\ g_2 \text{ true}\} \\ &+ Pr\{\text{rejection of second, } g_2\,|\,g_1 \ \&\ g_2 \text{ true}\} \\ &- Pr\{\text{rejection of } g_1 \text{ and } g_2\,|\,g_1 \ \&\ g_2 \text{ true}\} \end{aligned}$$

It seems appropriate to assume that the third term of the RHS is (approximately) zero since that joint event is unlikely. Hence

$$\alpha = \alpha_1 + \alpha_2.$$

Thus with $\alpha = .05$, it would seem logical to let $\alpha_1 = \alpha_2 = \alpha/2$. This is further supported by the observation that the expected numbers of positives and negatives are equal when $\rho \rightarrow 1$ since

$$Pr\{Y > 0\} = \lambda/(\lambda + \mu) = \rho/(\rho + 1) \rightarrow 1/2.$$

Later on in the paper, it will be possible for $\rho << 1$, in which case the numbers of positives would not equal that of the negatives. It is still true that the errors will be additive, but the size of the sub-samples will greatly affect the width of the acceptance region. Nevertheless, both sub-hypotheses must be accepted for the global one to be true, so that the acceptance region is defined in R^2 space as

$$\Big[(x,y)\,\Big|\, a_1 < x < b_1,\ a_2 < y < b_2, \ \text{ and}$$
$$Pr\,\{a_1 < x < b_1\} = Pr\{a_2 < y < b_2\} = 1-\alpha/2\Big].$$

[1] Note that y is zero with probability equal to zero; we shall not thus include such a possibility in the subsequent probability calculations.

[2] *It should be recognized that y might be zero in a real problem if the computations are not carried out to any great level of accuracy. In such cases it is probably best to assign the zero randomly to either of the two groups.*

Note that we proceed through the entire test without any knowledge whatsoever of the actual values of λ and μ. If, however, values of λ and μ are required, then their maximum-likelihood estimates are given by

$$\begin{cases} 1/\hat{\lambda} = \dfrac{\text{Sum of positive } y\text{'s}}{\text{Number of positive } y\text{'s}} & (10) \\ 1/\hat{\mu} = \dfrac{\text{Sum of negative } y\text{'s}}{\text{Number of negative } y\text{'s}} & (11) \end{cases}$$

On the other hand, suppose instead that μ and/or λ is specified. The unspecified parameter (if there is one) is handled exactly as in the preceding consistent with Equations (8) and (9). But if (for example) μ is known, then Equation (8) says that the positive values of y should follow a *known* exponential distribution. Hence their mean is Erlang distributed and we can then construct a suitable rejection region for the problem. This latter test should be done with some caution since its power is very much dependent upon the class of possible alternative hypotheses. This approach is uniformly most powerful against another exponential as alternative, but might have to be modified against, for example, a gamma alternative.

Of course, there have been other tests posed for exponentiality, inluding those working from the empirical CDF (exemplified in the work of Durbin [2]), rank-type tests (proposed originally by Proschan and Pyke [10]), and others built upon special characterizations of the exponential culminating in the recent work of Wang and Chang [12]). However, we stayed with the ratio test in view of its very desirable properties as noted in [9], namely, good power results when handled properly, ability to handle all levels of censoring, ease of computation, etc.

CASE B: Now let us discuss the case in which some of the line waits $\{W_q^{(n)}, n=1,2, \ldots\}$ could be zero. When a $W_q^{(n)}$ is indeed zero, it means that the associated customer goes straight into service and that the service facility had been idle from the time of departure of the previous customer (at least) until his arrival. This implies that the first difference $[W_q^{(n+1)} - W_q^{(n)}]$ is bounded below by $[S^{(n)} - T^{(n)}]$, that is,

$$W_q^{(n+1)} - W_q^{(n)} \geqslant S^{(n)} - T^{(n)}, \tag{12}$$

and furthermore that we may write

$$S^{(n)} - T^{(n)} = W_q^{(n+1)} - W_q^{(n)} - I^{(n)}, \tag{13}$$

where $I^{(n)}$ is the time ($\geqslant 0$) for which the server is idle within the nth waiting epoch.

The difficulty here is that every time the server is idle, we do not know the exact duration of its idle time, though we do have the number of $\{W_q^{(n)}\}$ which are zero. But at least we can find some related limits and expectations. Since the queue is M/G/1, successive idle periods are IID exponentials with the same mean λ^{-1} as that of the interarrival times. Thus λ may be estimated (call it $\tilde{\lambda}$) by the observed input rate[1]. We then recommend a heuristic alteration of the fundamental test statistic whereby $\tilde{\lambda}^{-1}$ is used for each $I^{(n)} > 0$. So in the event that $W_q^{(n+1)} = 0$ we have

[1] If this is not available, then we suggest the following approach. Consider the percentage of $\{W_q^{(n)}\}$ equal to zero as an estimator of $1-\rho$ (say $1-\bar{\rho}$) and equate it to $1-\tilde{\lambda}/\tilde{\mu}$. Then take the average line wait (say $\bar{W}_q$) and set that equal to $\tilde{\lambda}/(\tilde{\mu}(\tilde{\mu}-\tilde{\lambda}))$. The simultaneous solution of these would provide the necessary value of $\tilde{\lambda}$, namely, $\tilde{\lambda} = \bar{\rho}^2/[\bar{W}_q(1-\bar{\rho})]$.

$$S^{(n)} - T^{(n)} \doteq W_q^{(n+1)} - W_q^{(n)} - 1/\tilde{\lambda} \tag{14}$$

$$\doteq W_q^{(n)} - 1/\tilde{\lambda}.$$

Therefore, given the line waits from the M/G/1 queue, we can find the values of the $[S^{(n)} - T^{(n)}]$ by using either Equation (3) or Equation (14), depending upon whether $W_q^{(n+1)}$ is positive or zero, and then carry out the test of hypothesis on the positives and negatives as described before. It is important to use two distinct estimators of λ (say $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$) for the numerator and denominator of the test statistic, respectively, with each computed from the appropriate portion of the data.

This is clearly an approximate procedure though it is true that the average of times calculated for both the numerator and denominator of the Q ratio formed according to Equation (1) will indeed converge to $1/\lambda$ rather rapidly because the actual sequence $\{I^{(n)}\}$ obeys the strong law of large numbers for IID exponentials. To assess the quality of the approximation we ran some limited Monte Carlo tests for a few stationary queues using moderate to high values of the traffic intensity ρ and with varying sample sizes. Some results are presented in the following table for one such set of runs. For this we made 100 runs of an M/M/1 simulator with 100 customers, and $1/\lambda = 120$ and $1/\mu = 90$. For each run we computed the values of $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ and used them in the computation of the test statistics wherever appropriate. Since the number of negative values of $[S^{(n)} - T^{(n)}]$ must greatly exceed the number of positives, we worked only with the negatives and tested them for exponentiality.

TABLE 1. *Comparison of Empirical Simulation Ratio Statistic to F Critical Values ($\rho = 3/4$, N=100)*

	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10
Simulation	.508	.515	.556	.576	.598	.614	.637	.646	.671	.683
$F_{50,50}$	.514	.555	.584	.606	.625	.642	.656	.670	.682	.693
Approx. 95% Confidence Interval for Simulation %-iles	--	--	--	--	(.508, .671)	(.515, .683)	(.515, .696)	(.556, .702)	(.556, .715)	(.576, .728)

From the theory of order statistics we were able to construct confidence intervals from repeated runs for the 5% or 95% points (or any other, one at a time) of the empirical ratio statistic. It turned out that the comparable F-distribution critical points fell within these intervals, and thus we conclude that it is statistically legitimate to use the F-distribution for any related testing problem. When the sample size is fixed, the effectiveness of the approximation would improve as ρ gets closer to 1.

The key result which permits this to be accomplished is the fact that the order statistics of a random sample from $U(0,1)$ each have a beta distribution and that the joint CDF of any two is an ordered bivariate Dirichlet distribution. Thus the confidence coefficient of any confidence interval for an arbitrary percentile for the empirical ratio statistic may be computed using incomplete beta functions where the incomplete beta function may then be converted easily into the more easily handled binomial sum.

The System-Wait Problem (II)

The second problem that was considered is similar to the first one, except that we are given the system waiting time of each customer of an M/G/1 queue as opposed to the line waits of the previous problem. Again, these system waits are correlated to each other and so they are not easy to handle. The problem is more difficult if we do not know anything about the corresponding line waits.

CASE A: We first discuss the simplest case in which, in addition to system waits of each customer, we are also given to understand that none of the corresponding line waits is zero. For any GI/G/1 queue, if $W^{(n)}$ and $W_q^{(n)}$ are the system and line waits of the n th customer, respectively, then

$$W^{(n)} = W_q^{(n)} + S^{(n)} \tag{15}$$

where $S^{(n)}$ is the service time of the n th customer. Therefore

$$W^{(n+1)} - W^{(n)} = W_q^{(n+1)} - W_q^{(n)} + S^{(n+1)} - S^{(n)}. \tag{16}$$

Since no $W_q^{(n)}$ is zero by assumption, Equation (16) can be rewritten using Equation (3) as

$$\begin{aligned} W^{(n+1)} - W^{(n-1)} &= S^{(n)} - T^{(n)} + S^{(n+1)} - S^{(n)} \\ &= S^{(n+1)} - T^{(n)}. \end{aligned} \tag{17}$$

Since $T^{(n)}$ and $T^{(n-1)}$ are IID, the distribution of $[S^{(n+1)} - T^{(n)}]$ has already been derived and is given by Equation (5). We can now test the hypothesis that $G = M$ from the system waits of the M/G/1 queue, by testing whether the first differences of the system waits have come from the distribution given by (5).

CASE B: Suppose instead of assuming that none of the corresponding line waits is zero, we are told that some of them are indeed zero. That is, we are given the system waits of an M/G/1 queue with the additional knowledge as to which of the corresponding line waits are zero, but not the exact values of other corresponding line waits. Equation (16) is still valid, and by substitution for $[W_q^{(n+1)} - W_q^{(n)}]$ in Equation (14), we have

$$\begin{aligned} W^{(n+1)} - W^{(n)} &= S^{(n+1)} - T^{(n)} + I^{(n)} + S^{(n+1)} - S^{(n)} \\ &\doteq S^{(n)} - T^{(n-1)} + 1/\tilde{\lambda}, \end{aligned} \tag{18}$$

or

$$S^{(n+1)} - T^{(n)} \doteq W^{(n+1)} - W^{(n+1)} - 1/\tilde{\lambda}, \tag{19}$$

which is similar to Equation (14). Thus we can find the values of $[S^{(n+1)} - T^{(n)}]$ using Equation (19) and then proceed with the test of hypothesis $G = M$ as before.

CASE C: Suppose now that we are given the system waits for the M/G/1 queue, but do not known anything about the corresponding line waits. The solution of the problem is obtained using some basic concepts of probability for the case of testing when a parameter λ or μ is specified and using a Monte Carlo technique if neither is specified.

For the first subcase, suppose we are given the system waits of each customer, $W^{(1)}$, $W^{(2)}, \ldots, W^{(n)}$, but do not know which of the corresponding $W_q^{(n)}$ are zero, that is, when the system is empty. We now use an argument similar to that of Case I-B again by noting that the limiting expectation of $W^{(n)}$ is $1/(\mu-\lambda)$. Thus we may estimate $P_0 = 1-\lambda/\mu$, and assuming for example that μ is given, λ would be determined as indicated. We would then use Equation (19) as before.

Now if, instead, there is no knowledge about λ and μ the problem becomes more complex.

The key result which is employed in arriving at a reasonable solution to the problem is the conditional distribution of $T^{(n-1)}$ given the values of $W^{(n)}$. If we define

$$\nabla W^{(n)} \equiv W^{(n+1)} - W^{(n)} = S^{(n+1)} - T^{(n)},$$

then the conditional density of $T^{(n)}$ given $[\nabla W^{(n)} = d]$ is

$$f(t|d) = \frac{f_1(t)\, f_2(d+t)}{g(d)}.$$

From Equation (5) this density can be found as

$$f(t|d) = \begin{cases} \dfrac{\lambda e^{-\lambda t}\mu e^{-\mu(d+t)}}{[\lambda\mu/(\lambda+\mu)]e^{d\lambda}} & \begin{bmatrix} d < 0, \\ t > -d \end{bmatrix} \\[2ex] \dfrac{\lambda e^{-\lambda t}\mu e^{-\mu(d+t)}}{[\lambda\mu/(\lambda+\mu)]e^{-d\mu}} & (d > 0) \end{cases} = \begin{cases} (\lambda+\mu)\ e^{-(\lambda+\mu)(t+d)} & \begin{bmatrix} d < 0, \\ t > -d \end{bmatrix} \\[2ex] (\lambda+\mu)\ e^{-(\lambda+\mu)t} & (d>0) \end{cases}$$

Thus we see that the conditional distribution of the interarrival times is either exponential or shifted exponential according to whether $d > 0$ or $d < 0$. We are almost in a position to construct another positive/negative exponential test. We say almost for two reasons: (1) the values of t corresponding to negative d have a location parameter (this is no problem — the test can be adapted); (2) the $[\nabla W^{(n)}]$ are not IID (this poses a problem — but we shall assume that the resultant t values do form a (nearly) random sample). We need to do some Monte Carlo testing for this approach but our earlier experiences suggest success.

ISSUES FOR FURTHER WORK AND CONCLUDING REMARKS

There are a few places in this study where further investigations would be valuable and could tighten up the procedures somewhat. As examples of the kinds of issues which might be worth pursuing, we suggest the following: (1) further exploration of the implications of our approach to the Type I error when the queue traffic is indeed quite low; (2) to derive comparative power results for feasible alternative hypotheses such as mixed exponentials; and (3) to study the nature of the approximations $\bar{\lambda}_1$ and $\bar{\lambda}_2$ used in Part I, Case B. In fact, one referee has suggested that we model the situation in a decision-theoretic set-up to arrive at a decision rule, that minimizes some sort of risk function which incorporates any sample size effects. This is an interesting possibility for some new research.

An interesting additional issue is raised if instead of having an M/G/1 queue, we have GI/M/1 and are given line waits and/or system waits and are asked to test for exponentiality of the interarrival distribution; the entire analysis would be similar and we can carry out the test as before. This is true because Equations (2)-(9) are all still valid. Furthermore, by virtue of the continuity of basic GI/G/1 queue properties (for example, see [13]), the size of the test is likely to be approximately equal to the stated significance level even if the underlying arrival process is non-Poisson. However, the extension to more complex GI/G/1 systems is a complicated matter, but certainly a possibility in some specific queues and should be handled on a case-by-case basis.

BIBLIOGRAPHY

[1] Barlow, R.E., and F. Proschan, "A Note on Tests for Monotone Failure Rate Based on Incomplete Data," Annals of Mathematical Statistics **40**, 595-600.

[2] Durbin, J. "Kolmogorov-Smirnov Tests When Parameters are Estimated with Applications to Tests of Exponentiality and Tests on Spacings," Biometrika **62**, 5-22 (1975).

[3] Fercho, W.W., and L.J. RINGER, "Small Sample Power of Some Tests of Constant Failure Rate" Technometrics **14**, 713-724 (1972).

[4] Gnedenko, B.V., Y.K. Belyayev, and A.D. Solovyev, *Mathematical Methods of Reliability Theory,* (Academic Press, New York 1969).

[5] Gross, D., and C.M. Harris, "On One-For-One Ordering Inventory Models With State-Dependent Leadtimes," Operations Research **19**, 735-760 (1971).

[6] Gross, D., and C.M. Harris, "Continuous Review (*s,S*) Inventory Models With State-Dependent Leadtimes," Management Science **19**, 567-574 (1973).

[7] Gross, D., C.M. Harris, and J.A. Lechner, "Stochastic Inventory Models With Bulk Demand and State-Dependent Leadtimes," Journal of Applied Probability **8**, 521-534 (1971).

[8] Hadley, G., and T.M. Whitin, *Analysis of Inventory Systems,* (Prentice-Hall, Englewood Cliffs, New Jersey 1963).

[9] Harris, C.M., "A Note on Testing for Exponentiality," Naval Research Logistics Quarterly, **23**, 169-175 (1976).

[10] Proschan, F. and R. Pyke, "Tests for Monotone Failure Rate," Fifth Berkely Symposium on Mathematical Statistics and Probabilities 293-312 (1967).

[11] Stollmack, S., and C.M. Harris, "Failure-Rate Analysis Applied to Recidivism data," Operations Research **22**, 1192-1205 (1974).

[12] Wang, Y.H., and S.A. Chang, "A New Approach to the Nonparametric Tests of Exponential Distribution With Unknown Parameters," in *The Theory and Applications of Reliability,* (Academic Press, New York 1977).

[13] Whitt, W., "The Continuity of Queues," Advances in Applied Probability **6** 175-183 (1974).

SCHEDULING JOBS TO TWO MACHINES SUBJECT TO BATCH ARRIVAL ORDERING

Arie Tamir

Department of Statistics
Tel Aviv University
Tel Aviv, Israel

ABSTRACT

The problem considered is to assign n jobs to two processors so as to minimize the total flow time, with the constraint that a predetermined partial ordering (induced by batch arrivals) must be preserved within the subset of jobs assigned to each processor. An efficient algorithm of time $0(n^5)$ is developed, and computational experience is reported.

INTRODUCTION

This paper considers a problem of assigning n jobs to two parallel processors (machines). It is assumed that the jobs have a predetermined partial ordering, reflecting the order of batch arrivals. The processing times of the different jobs by the two (not necessarily identical) processors are known, and each job has to be processed by either one of the two processors. The objective is to assign the jobs to processors so as to minimize the total flow time, with the constraint that the original partial ordering must be preserved within the subset of jobs assigned to each machine. The case when the partial ordering is complete and the two processors are identical was solved in [7], using an efficient (polynomially bounded) dynamic programming approach. An extension of that situation to the case where processing times may depend on the processor as well as on the job were recently presented in [8,9]. Turning to a second extreme case, i.e. when the ordering is empty, we observe that the above problem can be solved efficiently using the formulation of [1,5]. In our setting the empty ordering will correspond to a joint arrival of all the jobs, while the complete ordering will correspond to the case of no simultaneous arrivals.

To solve our model we combine the dynamic programming approach of [7] with the assignment problem formulation of [1,5] to yield an efficient algorithm, whose time complexity is $0(n^5)$. Computational experience is provided in the last section.

THE MODEL

Consider a service center consisting of a waiting facility and a service department. The center operates as follows: customers may enter the waiting facility as long as it remains open; a period during which the service department is closed. Then the waiting facility is closed and additional customers are rejected. At this time the service department begins to serve customers which are already in the waiting facility. Each customer (job) can be served (processed) on either one of two available parallel machines, which are not necessarily identical. However, the processing of a job cannot be interrupted, once it has been started, nor can it be divided between the two processors. One cycle of operation ends after all the customers who are in the waiting facility have been completely serviced. At that point in time the service department closes, the waiting facility opens and new customers are admitted.

The nature of the system allows for batch arrivals as well as single arrivals, of customers to the waiting facility. The order of arrivals induces the following priority constraints on the scheduling of the jobs on the two machines. If two jobs have not arrived in the same batch and both are assigned to be processed by the same machine, then the one who joined the waiting facility earlier is to be processed first. (Note that this constraint does not apply to jobs arriving in the same batch).

Define the flow time of a job to be the time that elapses between the minute the service department opens and the completion of that job. Our objective is to assign and schedule the jobs in a way that minimizes the total (or average) flow over all jobs in the present cycle, subject to the above priority constraint.

Suppose that n jobs $J_1, J_2, \ldots, J_n$ are available for processing when the waiting facility closes. We say that job J_u "precedes" job J_t $(J_u > J_t)$ if and only if J_u has arrived before J_t. (If both jobs arrived in the same batch then neither $J_u > J_t$ nor $J_t > J_u$ and we write $J_u \approx J_t$). Denoting by t_{ij}, $i=1, 2, j=1, \ldots, n$, the (deterministic) processing time of job J_j by processor i, the objective is to minimize the total flow subject to the constraint that if $J_u > J_t$ and both jobs are assigned to the same machine, then J_u should be processed first.

At this point we observe that for the model considered in [7,8,9] we have either $J_t > J_u$ or $J_u > J_t$ but not $J_u \approx J_t$, while the model discussed in [1,5] satisfies $J_u \approx J_t$ for all pairs of jobs J_u and J_t. Equivalently, in our setting the first model corresponds to the case of no batch arrivals while the latter corresponds to a joint arrival of all the jobs.

Assuming that the n jobs have arrived in k-batches, we use t_p to denote the number of jobs in the p^{th} batch and label every one of them as a p-job. (We shall also say that the job is of type p). Note that $t_p \geq 1$, $p=1,2,\ldots,k$ and $t_1 + t_2 + \ldots + t_k = n$. We also assume that batches are numbered according to their order of arrival. T_p is defined by $T_p = t_p + t_{p+1} + \ldots + t_k$.

Since our objective is to minimize total flow (subject to a certain ordering) it is worthwhile to note a basic property of that criterion. Suppose that jobs $J_{j(1)}, \ldots, J_{j(r)}$ are assigned to machine i and assume further that $J_{j(m)}$ precedes $J_{j(m+1)}$, for $m = 1, \ldots, r-1$. Then the total flow of these jobs is $\sum_{m=1} (r-m+1) t_{ij(m)}$. Thus, the contribution of the processing time of job $J_{j(m)}$ depends only on the number of jobs it precedes but not on their processing times. For future reference we call this property the Basic Flow Property (BFP).

Suppose that the minimum flow is achieved when m_p of the t_p p-jobs, $p=1, \ldots, k$, are allocated to processor 1. This is illustrated in the next figure.

Processor 1 ←	m_1	. .	m_p	. . .	m_k
Processor 2 ←	$t_1 - m_1$	. .	$t_p - m_p$	. . .	$t_k - m_k$

Given, p, $1 \leq p \leq k$, the BFP then implies that the contribution of processing times of all jobs of types greater than or equal to p can be obtained by ignoring jobs of type less than p. Hence, suppose that only j-jobs, $p \leq j \leq k$ are available for processing and denote by $g_p(1,m)$, $m=0,1,\ldots,T_p$, the minimum flow of all jobs of type greater than or equal to p, provided m of them are processed by machine 1.

To find $g_p(1,m)$ assume the number of type p jobs assigned to machine 1 is m_p, $0 \leq m_p \leq m$. (Note that m_p has also to satisfy $m_p \leq t_p$ and $m-m_p \leq T_{p+1}$). Let $f_p(1,m_p,m)$ denote the total contribution of the processing times of p-jobs to the total flow $g_p(1,m)$, when m_p p-jobs are assigned to machine 1. By the BFP it follows that

$$g_p(1,m) = \underset{m_p}{\text{minimum}} \{f_p(1,m_p,m) + g_{p+1}(1,m-m_p)\} \tag{1}$$

$p=1,\ldots,k-1$, where m_p varies between $\max(0,m-T_{p+1})$ and $\min(m,t_p)$.

To solve our original sequencing problem we use the recursive relation (1) to compute $g_1(1,m)$ for all integers m such that $0 \leq m \leq T_1 = n$. The optimal solution is the minimum of $\{g_1(1,m) | m = 0,1,\ldots,n\}$.

To apply the recursive relation (1) we next show how to compute $f_p(1,m_p,m)$, $p=1,\ldots,k-1$, and $g_k(1,m-m_p)$ by using a linear assignment problem formulation. Our formulation is a modification of that in [1,5].

Let J_r be an arbitrary p-job. If J_r is assigned to machine i and it precedes $(u-1)$ p-jobs and v jobs of higher types on machine i, then the contribution of its processing time, t_{ir}, to the total flow of all the jobs is $(u+v)t_{ir}$. J_r is then said to have occupied the uth position at machine i. This position is denoted by (i,u). To find $f_p(1,m_p,m)$ we consider only those schedules that assign exactly m_p p-jobs and $m-m_p$ jobs of higher types to machine 1. Define a linear assignment problem by the following bipartite graph. The first set of nodes consisting of t_p nodes, corresponds to the t_p p-jobs. A typical node is denoted by (J_r) where J_r is a p-job. The other set of nodes is the set of t_p positions assigned to the p-jobs, i.e., m_p positions on machine 1 and t_p-m_p positions on machine 2. These nodes are labelled as (i,u) nodes, $i=1,2$. In the corresponding bipartite graph each (J_r) node is connected to each (i,u) node. Furthermore if (J_r) is assigned $(1,u)$, $u=1,\ldots,m_p$, then the 'cost' associated is $(u+m-m_p)\, t_{1r}$ since $m-m_p$ jobs of type higher than p are also assigned to machine 1. Similarly an assignment of (J_r) to $(2,u)$, $u=1\ldots,t_p-m_p$, yields a cost of $(u+T_{p+1}-(m-m_p))t_{2r}$ where $T_{p+1} = t_{p+1} + t_{p+2} + \ldots + t_k$.

Following arguments similar to those in [1,5] we easily verify that the minimum solution of the above linear assignment problem is indeed equal to $f_p(1,m_p,m)$. (We call the above assignment the (p,m,m_p) assignment). We also realize that an almost identical scheme computes $g_k(1,m)$ for $m=0,1,\ldots,t_k$. If J_r is a k-job which is assigned to position $(1,u)$, $u=1,\ldots,m$, (or $(2,u)$, $u=1,\ldots,t_k-m$), then it is followed on machine 1 (machine 2), by exactly $(u-1)$ jobs of its type and no jobs of other types. Thus we may suggest the above assignment formulation for the solution of $g_k(1,m)$, $m=0,1,\ldots,t_k$. The 'cost' associated with the arc connecting (J_r) and (i,u), $i=1,2$, is ut_{ir}.

We summarize our solution approach with the following scheme.

STEP 1 For each $m=0,\ldots,t_k$ compute $g_k(1,m)$. Set $p=k-1$, and $m=0$.

STEP 2 For any integer m_p satisfying $\max(0,m-T_{p+1}) \leq m_p \leq \min(m,t_p)$ compute $f_p(1,m_p,m)$, the minimum of the (p,m,m_p) assignment.

STEP 3 Define

$$g_p(1,m) = \underset{m_p}{\text{minimum}}\{f_p(1,m_p,m) + g_{p+1}(1,m-m_p)\}.$$

If $m = T_p$ go to Step 4, otherwise increase m by 1 and go to Step 2.

STEP 4 Decrease p by 1. If $p = 0$ go to Step 5. Otherwise, set $m = 0$ and go to Step 2.

STEP 5 Find the minimum of $\{g_1(1,0), \ldots, g_1(1,n)\}$.

Several comments are in order. When one considers the case of [7,8,9] i.e. no batch arrivals, then the corresponding (p,m,m_p) assignments are trivial since m_p' takes on the values 0,1 only.

Let J_p denote the only p-job. Then it is easily verified that $g_k(1,1) = t_{1k}$ and $g_k(1,0) = t_{2k}$. (Note that $k = n$). Similarly, we obtain $f_p(1,1,m) = mt_{1p}$ and $f_p(1,0,m) = (T_p-m)t_{2p} = (n+1-p-m)t_{2p}$ for $m = 1,2,\ldots,n-p$. If $m = 0$ then we have $f_p(1,0,0) = T_p \cdot t_{2p} = (n+1-p)t_{2p}$ while the case $m = T_p = (n-p+1)$ yields $f_p(1,1,n+1-p) = (n+1-p)t_{1p}$.

Thus, for $p = 1,\ldots,k-1$ the recursive relation (1) is replaced by

$$g_p(1,m) = \mathit{minimum}\ \{mt_{1p} + g_{p+1}(1,m-1);(n+1-p-m)t_{2p} + g_{p+1}(1,m)\}, \tag{2}$$

when $m = 1,2,\ldots,n-p$, by $g_p(1,m) = (n+1-p-m)t_{2p} + g_{p+1}(1,m)$ when $m = 0$, and by $g_p(1,m) = mt_{1p} + g_{p+1}(1,m-1)$ when $m = n-p+1$.

Turning to the other extreme case, i.e., all jobs arrive at the same time, then $k=1$ and it suffices to solve $(n+1)$ assignment problems (each with $2n$ nodes) to obtain minimum $\{g_1(1,0), g_1(1,1), \ldots, g_1(1,n)\}$. In fact a single assignment with $3n$ nodes can replace the above $(n+1)$ problems to yield the optimal solution. (See [1,5]).

Finally we point out that if processing times depend only on the job but not on the processor, i.e. the identical machine case, then batch arrivals can be reduced to the situation of single arrivals. This is done by ordering jobs of the same batch according to increasing processing times and then applying the dynamic programming algorithm of [7].

COMPUTATIONAL EFFORT

To find the total number of calculations required to solve the optimal policy by our method, we first focus on the (p,m,m_p) assignments.

Computing $g_k(1,m)$, $m = 0,1,\ldots,t_k$, involves the solution of (t_k+1) assignment problems, where each is associated with a bipartite graph of $2t_k$ nodes. In the next step $f_{k-1}(1,m_{k-1},m)$ is computed for all $m = 0,1,\ldots,T_{k-1}$ and $\max(0,m-T_k) \leq m_{k-1} \leq \min(m,t_{k-1})$. In general, to find $g_p(1,m)$, $p = 1,\ldots,k-1$; $m = 0,1,\ldots,T_p$, from $g_{p+1}(1,m)$, $m = 0,1,\ldots,T_{p+1}$, our scheme requires the computation of $f_p(1,m_p,m)$ for all $m = 0,1,\ldots,T_p$ and $\max(0,m-T_{p+1}) \leq m_p \leq \min(m,t_p)$. Partitioning the feasible domain of the indices (m_p,m) shows that the number of (p,m,m_p) assignment problems - each with $2t_p$ nodes - solved by the scheme is

$$A_p = (S_p - s_p)(s_p+1) + (s_p+1)(s_p+2)/2 + (T_p-S_p)(T_p-S_p+1)/2$$

where $S_p = \max(t_p, T_{p+1})$ and $s_p = \min(t_p, T_{p+1})$.

Following [3,4,6] we note that a solution of an assignment problem with $2t_p$ nodes requires $0(t_p^3)$ elementary operations. Hence the total computational effort spent on solving the (p,m,m_p) assignments is

$$0[(1+T_k)t_k^3 + \sum_{p=1}^{k-1} A_p \cdot t_p^3] = 0(n^2(n-k+1)^3)$$

From (1) it follows that in addition to the above computational effort we have to perform certain comparisons and additions. A simple analysis shows that the latter effort is bounded by $0(n^2)$. Hence the total number of calculations spent to solve the original sequencing problem is dominated by $0(n^2(n-k+1)^3)$, where n is the number of jobs and $k \leq n$ is the number of bulks.

The above algorithm has been programmed and tested on several problems. Using the CDC-6500 computer installation at Tel Aviv University we have run problems with randomly generated data of up to 300 jobs clustered into up to 20 batches. Early results yield execution times of 3 to 5 seconds for the 300-job problems. Since we have used an ordinary algorithm to solve the assignment problems, we feel that run times can be further decreased by implementing the algorithm reported in [2, Section 3.9], which is specially designed to solve the assignment problems arising in our model.

ACKNOWLEDGEMENT

I would like to thank Mrs. Sarah Rabinovitch for programming and testing the algorithm.

BIBLIOGRAPHY

[1] Bruno, J., E. G. Coffman and R. Sethi , "Scheduling Independent Tasks to Reduce Mean Finishing Time", Communications of the Associations for Computing Machinery, *17*, 382-387, (1974).

[2] Coffman, E. G., *Computer and Job/Shop Scheduling Theory*, (John Wiley, New York, 1976).

[3] Gabow, H. N., "An Efficient Implementation of Edmond's Algorithm for Maximum Matching on Graphs," Journal of the Association for Computing Machinery, *23* (1976).

[4] Hopcroft J. E. and R. M. Karp, " An $n^{5/2}$ Algorithm for Maximum Matchings in Bipartite Graphs," SIAM J. on Computing, *2*, 225-231 (1973).

[5] Horn W. A., "Minimizing Average Flow Time with Parallel Machines," Operations Research, *21*, 846-7 (1973).

[6] Kuhn, H. W., " The Hungarian Method for the Assignment Problem," Naval Research Logistics Quarterly, *2*, 382-387 (1955).

[7] Mehta, S. R., R. Chandrasekaran and H. Emmons, "Order-Preserving Allocation of Jobs to Two Machines," Naval Research Logistics Quarterly, *21*, 361-364 (1974).

[8] Rothkopf, M. H., "A Note on Allocating Jobs to Two Machines," Naval Research Logistics Quarterly, *22*, 829-830 (1975).

[9] Zaloom V. and D. Vatz, "A Note on the Optimal Scheduling of Two Parallel Processors," Naval Research Logistics Quarterly, *22*, 823-827 (1975).

SCHEDULING OF STOCHASTIC TASKS ON TWO PARALLEL PROCESSORS*

Michael Pinedo

Instituto Venezolano de Investigaciones Cientificas
Caracas, Venezuela

Gideon Weiss

Department of Statistics
Tel-Aviv University
Tel-Aviv, Israel

ABSTRACT

We consider the problem of scheduling n tasks on two identical parallel processors. We show both in the case when the processing times for the n tasks are independent exponential random variables, and when they are independent hyperexponentials which are mixtures of two fixed exponentials, that the policy of performing tasks with longest expected processing time (LEPT) first minimizes the expected makespan, and that in the hyperexponential case the policy of performing tasks with shortest expected processing time (SEPT) first minimizes the expected flow time. The approach is simpler than the dynamic programming approach recently employed by Bruno and Downey.

1. INTRODUCTION

Suppose that two identical parallel processors are available to perform tasks 1, ..., n with random processing times $X_1, \dots, X_n$. Performing task j requires one of the processors (either of the two can do) for a duration X_j, after which it is complete. Tasks are performed consecutively, starting at time $t = 0$, so that as soon as a task is complete another task is put on the processor that is freed. For any permutation $j_1, \dots, j_n$ of 1, ..., n, putting tasks on the processors in that order defines a schedule. It is assumed for every task j that X_j is independent of all other X_i's, of the schedule, of the moment its processing starts and of the processor on which it is performed.

Two criteria are commonly used to judge the performance of a schedule: *Flow time* — the sum over all tasks of the time from zero to completion, and *makespan* — the time from zero until all tasks are completed. These give rise to two problems:

PROBLEM I:

Find a schedule that minimizes the expected flow time.

*This work was done while the authors were at the University of California, Berkeley, at the Department of Industrial Engineering and Operations Research and the Department of Statistics.

PROBLEM 2:

Find a schedule that minimizes the expected makespan.

Two special schedules based on the expectations of $X_1, \ldots, X_n$ are intuitively appealing:

SEPT - (shortest expected processing time first) which uses the order $j_1, \ldots, j_n$ where $E(X_{j_1}) \leqslant E(X_{j_2}) \leqslant \ldots \leqslant E(X_{j_n})$; and

LEPT - (longest expected processing time first) which uses the order $j_n, \ldots, j_1$.

In the case where $X_1, \ldots, X_n$ are deterministic (i.e., given real numbers), it is known that SEPT is the solution to Problem 1 [2]. However, LEPT does not in general minimize the makespan, and it was shown by Karp [3] that finding a schedule to solve Problem 2 is an NP-complete problem. By that one means that it belongs to a large collection of problems which have the same degree of computational complexity, and for which it is conjectured that no polynomial time algorithm exists to solve them. In the makespan problem this conjecture means that virtually all 2^n possible schedules need to be enumerated.

In the case where $X_1, \ldots, X_n$ are exponentially distributed, Bruno and Downey [1] have recently shown, using dynamic programming, that the SEPT and LEPT schedules do indeed solve Problems 1 and 2 respectively. We present here another proof that LEPT solves Problem 2.

We also consider an additional case, where each of $X_1, \ldots, X_n$ is distributed as a mixture of two fixed exponential random variables. In the context of application, this corresponds to the case where the work required by each task is either long, exponential with parameter μ_1 or short, exponential with parameter μ_2 ($\mu_1 < \mu_2$), and we know for each task the probability θ that it is long. In this special hyperexponential case we show that SEPT and LEPT solve problems 1 and 2 respectively. The approach is simpler than the dynamic programming approach of Bruno and Downey.

2. PRELIMINARIES

To the tasks $1, \ldots, n$, with associated durations $X_1, \ldots, X_n$ we add task 0, with duration X_0. We assume that $X_0, X_1, \ldots, X_n$ are independent nonnegative random variables. For a permutation $j_1, \ldots, j_n$ of $1, \ldots, n$ let $Y_1, \ldots, Y_n = X_{j_1}, \ldots, X_{j_n}$. We call $X_0, Y_1, \ldots, Y_n$ a schedule. The performance of the schedule proceeds from $t = 0$ as follows: at time $t = 0$ the previously scheduled task 0 is on one of the processors, and occupies it until X_0, when it is complete. Task j_1 is put on the other processor at $t = 0$, which it occupies until time Y_1, when it is complete. At the first completion, task j_2 is put on the freed processor, and tasks continue to be put on processors when they become available. We denote by:

$$(1) \qquad 0 \leqslant T_0(X_0, Y_1, \ldots, Y_n) \leqslant T_1(X_0, Y_1, \ldots, Y_n) \leqslant \ldots \leqslant T_n(X_0, Y_1, \ldots, Y_n)$$

the completion times of the tasks in their order of occurrence. Task j_{k+2} is put on a processor at $T_k(X_0, Y_1, \ldots, Y_n)$, $k = 0, 1, \ldots, n-2$, and occupies it until its completion at $T_k(X_0, Y_1, \ldots, Y_n) + Y_{k+2}$. The two processors complete their tasks at $T_{n-1}(X_0, Y_1, \ldots, Y_n)$ and $T_n(X_0, Y_1, \ldots, Y_n)$. We define

$$(2) \qquad D(X_0, Y_1, \ldots, Y_n) = T_n(X_0, Y_1, \ldots, Y_n) - T_{n-1}(X_0, Y_1, \ldots, Y_n).$$

The makespan of the schedule is $T_n(X_0, Y_1, \ldots, Y_n)$. Let

$$S = X_0 + X_1 + \ldots + X_n.$$

We note that:

$$\begin{aligned} S &= T_{n-1}(X_0, Y_1, \ldots, Y_n) + T_n(X_0, Y_1, \ldots, Y_n) \\ &= 2T_n(X_0, Y_1, \ldots, Y_n) - D(X_0, Y_1, \ldots, Y_n) \\ &= 2T_{n-1}(X_0, Y_1, \ldots, Y_n) + D(X_0, Y_1, \ldots, Y_n) \end{aligned} \tag{3}$$

and, since S is independent of the order in which tasks are performed, a schedule which minimizes the expected value of D will at the same time minimize the expected makespan. We shall prove the optimality of LEPT for Problem 2 by showing that it minimizes D stochastically.

The flow time of the schedule is the sum of all the completion times, $\sum_{k=0}^{n} T_k (X_0, Y_1, \ldots, Y_n)$. This differs from the sum of all the starting times $\sum_{k=0}^{n-2} T_k(X_0, Y_1, \ldots, Y_n)$ by S, so that the expectations of these two sums are minimized simultaneously. We shall show in the hyperexponential case that SEPT minimizes the expectation of each of $T_k(X_0, Y_1, \ldots, Y_n)$, $k = 0, 1, \ldots, n-2$.

If we look at the performance of tasks $0, j_1, \ldots, j_k$ $1 \leqslant k \leqslant n$ alone, and use definitions (1), (2) for k instead of n, we note that:

$$T_i(X_0, Y_1, \ldots, Y_n) = T_i(X_0, Y_1, \ldots, Y_k) \quad i = 0, 1, \ldots, k-1 \tag{4}$$

while

$$T_k(X_0, Y_1, \ldots, Y_n) = T_{k-1}(X_0, Y_1, \ldots, Y_k) + \min\,(Y_{k+1}, D(X_0, Y_1, \ldots, Y_k)). \tag{5}$$

We remark that the addition of task 0 generalizes the original problems which correspond to $X_0 = 0$. It disposes of the need to schedule two tasks simultaneously at $t = 0$, and thus makes the SEPT and LEPT schedules unique.

Notation:

$A \supseteq B$ denotes that A is stochastically greater than B, that is: $P(A > x) \geqslant P(B > x)$ for all x.

3. THE EXPONENTIAL CASE:

We now assume that $X_1, \ldots, X_n$ are exponentially distributed with rates $\mu_1, \ldots, \mu_n$. The next lemma examines the effect on D of changing a schedule by interchanging two consecutive tasks.

LEMMA 1:

For any X_n and for $\mu_1 = \min(\mu_1, \mu_2, \ldots, \mu_n)$

$$D(X_0, X_1, X_2, \ldots, X_n) \subseteq D(X_0, X_2, X_1, \ldots, X_n). \tag{6}$$

PROOF:

Denote by p_j (q_j), $j = 0, 1, \ldots, n$ the probability that the last task to be completed by the schedule $X_0, X_1, \ldots, X_n$ $(X_0, X_2, X_1, \ldots, X_n)$ is task j. We show by induction that

$$(7) \qquad \begin{aligned} p_0 &= q_0 \\ p_1 &\leqslant q_1 \\ p_j &\geqslant q_j \quad j = 2, \ldots, n. \end{aligned}$$

This will imply (6) since the distribution of $D(X_0, X_1, X_2, \ldots, X_n)$ (of $D(X_0, X_2, X_1, \ldots, X_n)$) is a mixture of the distributions of $X_0 - \sum_{j=1}^{n} X_j \mid X_0 > \sum_{j=1}^{n} X_j, X_1, \ldots, X_n$ with mixing probabilities $p_0, p_1, \ldots, p_n$ $(q_0, q_1, \ldots, q_n)$ and if (7) holds, the distribution of $D(X_0, X_2, X_1, \ldots, X_n)$ can also be regarded as a mixture of the distributions of $X_0 - \sum_{j=1}^{n} X_j \mid X_0 > \sum_{j=1}^{n} X_j, X_1, Z_2, \ldots, Z_n$ with probabilities $p_0, p_1, \ldots, p_n$, where Z_j is a mixture of X_1 and X_j with probabilities $(p_j - q_j)/p_j$, q_j/p_j, and since $X_j \subseteq X_1$, $X_j \subseteq Z_j$ $j = 2, \ldots, n$.

To check (7) for $n = 2$, condition on $X_0 = x$, to obtain:

$$p_0 = q_0 = P(X_1 + X_2 < x)$$

$$(8) \qquad p_1 = e^{-\mu_1 x} \frac{\mu_2}{\mu_1 + \mu_2} \qquad q_1 = 1 - q_0 - q_2$$

$$p_2 = 1 - p_0 - p_1 \qquad q_2 = e^{-\mu_2 x} \frac{\mu_1}{\mu_1 + \mu_2}$$

so:

$$(9) \qquad q_1 - p_1 = \frac{2\mu_1\mu_2}{\mu_2^2 - \mu_1^2} [e^{-\mu_1 x} - e^{-\mu_2 x}] \geqslant 0.$$

and get (7) by unconditioning.

For $n > 2$, denote by p'_j (q'_j) $j = 0, 1, \ldots, n-1$ the corresponding probabilities for $X_0, X_1, \ldots, X_{n-1}$ $(X_0, X_2, X_1, \ldots, X_{n-1})$, and assume (induction hypothesis) that:

$$(10) \qquad \begin{aligned} p'_0 &= q'_0 \\ p'_1 &\leqslant q'_1 \\ p'_j &\geqslant q'_j \quad j = 2, \ldots, n-1. \end{aligned}$$

Then:

$$(11) \qquad p_0 = q_0 = P(X_1 + X_2 + \ldots + X_n < X_0)$$

and

$$(12) \qquad p_j = p'_j \frac{\mu_n}{\mu_j + \mu_n}, \; q_j = q'_j \frac{\mu_n}{\mu_j + \mu_n} \qquad j = 1, \ldots, n-1.$$

So from (10)

(13)
$$p_1 \leqslant q_1$$
$$p_j \geqslant q_j \quad j = 2, \ldots, n-1.$$

Also, by $\mu_1 \leqslant \mu_j$ for all $j = 2, \ldots, n-1$,

(14)
$$\frac{\mu_n}{\mu_1 + \mu_n} \geqslant \frac{\mu_n}{\mu_j + \mu_n} \quad j = 2, \ldots, n-1$$

which with (10) implies

(15)
$$\sum_{j=1}^{n-1} p_j' \frac{\mu_n}{\mu_j + \mu_n} \leqslant \sum_{j=1}^{n-1} q_j' \frac{\mu_n}{\mu_j + \mu_n}$$

and so:

(16)
$$p_n \geqslant q_n. \blacksquare$$

We can now prove

THEOREM 1:

For arbitrary X_0 and for $X_1, \ldots, X_n$ independent exponential random variables, the LEPT schedule minimizes the expected makespan.

PROOF:

We assume $\mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_n$, so $E(X_1) \geqslant \ldots \geqslant E(X_n)$ and the LEPT schedule is $X_0, X_1, \ldots, X_n$. By (3) it is enough to show that for any schedule $X_0, Y_1, \ldots, Y_n$

(17)
$$E(D(X_0, X_1, \ldots, X_n)) \leqslant E(D(X_0, Y_1, \ldots, Y_n)).$$

We prove the stronger assertion, that

(18)
$$D(X_0, X_1, \ldots, X_n) \subseteq D(X_0, Y_1, \ldots, Y_n).$$

For $n = 2$, (18) holds by Lemma 1. Assume inductively that it holds for $n-1$, where $n > 2$. Look at an arbitrary schedule, $X_0, Y_2, \ldots, Y_k, X_1, Y_{k+1}, \ldots, Y_n$, where $k > 1$. We can regard $D(X_0, Y_2, \ldots, Y_k, X_1, Y_{k+1}, \ldots, Y_n)$ as $D(Z, Y_k, X_1, Y_{k+1}, \ldots, Y_n)$ where $Z = D(X_0, Y_2, \ldots, Y_{k-1})$. By Lemma 1 this is stochastically larger than $D(Z, X_1, Y_k, \ldots, Y_n)$ which is the same as $D(X_0, Y_2, \ldots, Y_{k-1}, X_1, Y_k, \ldots, Y_n)$. Apply this argument k times to get:

(19)
$$D(X_0, X_1, Y_2, \ldots, Y_n) \subseteq D(X_0, Y_2, \ldots, Y_k, X_1, Y_{k+1}, \ldots, Y_n).$$

Now regard $D(X_0, X_1, Y_2, \ldots, Y_n)$ as $D(Z', Y_2, \ldots, Y_n)$ where $Z' = D(X_0, X_1)$, and apply the induction hypothesis to obtain (18).

4. THE HYPEREXPONENTIAL CASE

We now assume that the distribution of X_j is $\bar{F}_j$, $j = 1, \ldots, n$ where

(20)
$$\bar{F}_j(x) = P(X_j > x) = \theta_j e^{-\mu_1 x} + (1 - \theta_j) e^{-\mu_2 x},$$

and $\mu_1 \leqslant \mu_2$. Obviously $X_i \supseteq X_j$ if $\theta_i \geqslant \theta_j$. The next lemma again examines the effect on D of interchanging two consecutive tasks.

LEMMA 2:

For arbitrary X_0, if $\theta_k \geqslant \theta_{k+1}$ for some $1 \leqslant k \leqslant n$, then:

$$D(X_0,X_1, \ldots, X_k,X_{k+1}, \ldots, X_n) \subseteq D(X_0,X_1, \ldots, X_{k+1},X_k, \ldots, X_n). \tag{21}$$

PROOF:

We consider first $k = 1$, so $\theta_1 \geqslant \theta_2$. We note that $D(X_0,X_1, \ldots, X_n)$ is either the remainder of task 0, $X_0 - \sum_{i=1}^{n} X_i \mid X_0 > \sum_{i=1}^{n} X_i$, or else it is an exponential random variable, either with rate μ_1 or with rate μ_2. We denote the probabilities of these three possibilities by $p_0 = P\left\{X_0 > \sum_{i=1}^{n} X_i\right\}$, p, $1 - p_0 - p$ respectively. We denote by q_0, q; p_0', p'; q_0', q' the probabilities of the same events for the schedules $X_0,X_2,X_1, \ldots, X_n$; $X_0,X_1, \ldots, X_{n-1}$; $X_0,X_2,X_1, \ldots, X_{n-1}$. Obviously $p_0 = q_0$, $p_0' = q_0'$, and to show that $D(X_0,X_1,X_2, \ldots, X_n) \subseteq D(X_0,X_2,X_1, \ldots, X_n)$ we have to prove that $p \leqslant q$.

For $n = 2$, condition on $X_0 = x$ to obtain,

$$p = \theta_1\theta_2(e^{-\mu_1 x} + \mu_1 x e^{-\mu_1 x}) + \theta_1(1 - \theta_2)e^{-\mu_1 x}\frac{\mu_2}{\mu_1 + \mu_2} + (1 - \theta_1)\,\theta_2\left\{e^{-\mu_2 x}\frac{\mu_2}{\mu_1 + \mu_2} + \frac{\mu_2}{\mu_2 - \mu_1}(e^{-\mu_1 x} - e^{-\mu_2 x})\right\} \tag{22}$$

and a similar expression for q. By subtracting:

$$q - p = (\theta_1 - \theta_2)\frac{2\mu_1\mu_2}{\mu_2^2 - \mu_1^2}(e^{-\mu_1 x} - e^{-\mu_2 x}) \geqslant 0, \tag{23}$$

which upon unconditioning shows that $q \geqslant p$.

Let now $n > 2$, and assume inductively that $q' \geqslant p'$. Then:

$$p = p'\theta_n + p'(1 - \theta_n)\frac{\mu_2}{\mu_1 + \mu_2} + (1 - p_0' - p')\,\theta_n\frac{\mu_2}{\mu_1 + \mu_2} + p_0'\theta_n P(X_1 + X_2 + \ldots + X_n > X_0 \mid X_1 + \ldots + X_{n-1} < X_0, X_n \sim \exp(\mu_1)) \tag{24}$$

where the last term on the RHS is independednt of the schedule of $X_1, \ldots, X_{n-1}$. A similar expression holds for q, and:

$$q - p = (q' - p')\left\{\theta_n\frac{\mu_1}{\mu_1 + \mu_2} + (1 - \theta_n)\frac{\mu_2}{\mu_1 + \mu_2}\right\} \geqslant 0. \tag{25}$$

For any $k > 1$, apply the lemma as proved for $k = 1$ to obtain that $D(Z,X_k,X_{k+1}, \ldots, X_n) \subseteq D(Z,X_{k+1},X_k, \ldots, X_n)$ which yields (21) when $Z = D(X_0,X_1, \ldots, X_{k-1})$.■

We can now prove:

THEOREM 2:

Let X_0 be arbitrary, $X_1, \ldots, X_n$ be hyperexponential random variables as in (20), with $\theta_1 \geqslant \theta_2 \ldots \geqslant \theta_n$, then:

(i) The LEPT schedule $X_0, X_1, \ldots, X_n$ minimizes the expected makespan.

(ii) The SEPT schedule $X_0, X_n, \ldots, X_1$ minimizes the expected flow time.

PROOF:

(i) Any schedule $X_0, Y_1, \ldots, Y_n$ can be changed by a sequence of steps, each involving an interchange between a longer task and a shorter task that directly precedes it, to the schedule $X_0, X_1, \ldots, X_n$. At each step the resulting random variable D decreases stochastically by Lemma 2. Part (i) follows by (3).

(ii) Let $X_0, Y_1, \ldots, Y_n$ be any schedule, and assume $Y_k \supseteq Y_{k+1}$. We compare this schedule with $X_0, Y_1, \ldots, Y_{k+1}, Y_k, \ldots, Y_n$. We note by (4) that:

(26)
$$T_i(X_0, Y_1, \ldots, Y_n) = T_i(X_0, Y_1, \ldots, Y_{k-1})$$
$$= T_i(X_0, Y_1, \ldots, Y_{k+1}, Y_k, \ldots, Y_n)$$
$$i = 0, 1, \ldots, k-2$$

and by (5) that for $k-1$:

$$T_{k-1}(X_0, Y_1, \ldots, Y_k, Y_{k+1}, \ldots, Y_n) =$$
$$T_{k-2}(X_0, Y_1, \ldots, Y_{k-1}) + \min(Y_k, D(X_0, Y_1, \ldots, Y_{k-1})) \supseteq$$

(27)
$$T_{k-2}(X_0, Y_1, \ldots, Y_{k-1}) + \min(Y_{k+1}, D(X_0, Y_1, \ldots, Y_{k-1})) =$$
$$T_{k-1}(X_0, Y_1, \ldots, Y_{k+1}, Y_k, \ldots, Y_n).$$

Finally we want to compare $T_i(X_0, Y_1, \ldots, Y_k, Y_{k+1}, \ldots, Y_n)$ with $T_i(X_0, Y_1, \ldots, Y_{k+1}, Y_k, \ldots, Y_n)$ for $i = k, k+1, \ldots, n-1$. From Lemma 2

(28)
$$D(X_0, Y_1, \ldots, Y_k, Y_{k+1}, \ldots, Y_{i+1}) \subseteq D(X_0, Y_1, \ldots, Y_{k+1}, Y_k, \ldots, Y_{i+1})$$

and this implies by (3)

(29)
$$E(T_i(X_0, Y_1, \ldots, Y_k, Y_{k+1}, \ldots, Y_{i+1})) \geqslant$$
$$E(T_i(X_0, Y_1, \ldots, Y_{k+1}, Y_k, \ldots, Y_{i+1}))$$

which by (4) is the same as:

(30)
$$E(T_i(X_0, Y_1, \ldots, Y_k, Y_{k+1}, \ldots, Y_n)) \geqslant$$
$$E(T_i(X_0, Y_1, \ldots, Y_{k+1}, Y_k, \ldots, Y_n)) \quad i = k, \ldots, n-1$$

Using (26), (27), (30) we have shown:

$$E\left\{\sum_{i=0}^{n-2} T_i(X_0, Y_1, \ldots, Y_k, Y_{k+1}, \ldots, Y_n)\right\} \geqslant$$

$$E\left\{\sum_{i=0}^{n-2} T_i(X_0, Y_1, \ldots, Y_{k+1}, Y_k, \ldots, Y_n)\right\}$$

That is, if $Y_k \supseteq Y_{k+1}$, their interchange will decrease the expected flow time. The proof is completed by an argument similar to Part (i).■

REMARKS:

The following remarks should be made with respect to our results.

(i) The policies that were shown to be optimal in Sections 3 and 4 are not only optimal in the class of $n!$ possible sequences which determine in advance (at time zero) which task will be next whenever one has been completed (independent of past history of the process), they are also optimal in that class of policies which allows the decision maker to review his policy at every task completion, taking into consideration what occurred before. To see why that is true assume that at the $n - k$ decision moment, when only k tasks are still to be scheduled, a schedule can be determined in advance for all the remaining decision moments. This is obviously true for $k = 1$. Theorems 1, 2 applied to the task currently processed as task 0, and the as yet unscheduled tasks as tasks $1, \ldots, k-1$ then show that the schedule must be LEPT for Problem 1 and SEPT for Problem 2. Hence at the $n - k - 1$ decision moment, the decision maker will choose the task to be scheduled immediately, and will also determine in advance the schedule of the k tasks not scheduled immediately.

We also note that in both cases discussed in Sections 3 and 4, when $X_0 = 0$, and LEPT is used, at any of the above decision moments the task already on the processor has the longest expected remaining processing time among all uncompleted tasks. This is trivially true for the exponential case, and is true for the hyperexponential case because the X_i's have DFR distributions. Hence a similar argument to the above shows that LEPT determined in advance is optimal among all the policies which allow rescheduling and preemption at decision moments.

(ii) The LEPT policy is also optimal when $X_1, \ldots, X_n$ are mixtures of exponentials with the rate of X_i a random variable Λ_i, and $P(\Lambda_1 < \Lambda_2 < \ldots < \Lambda_n) = 1$. However, if we only require $\Lambda_1 \subseteq \Lambda_2 \subseteq \ldots \subseteq \Lambda_n$, LEPT is not in general optimal as the following counterexample shows:

$$X_1 \sim \exp(1)$$

$$X_2 \sim \exp(\Lambda_1), \quad P(\Lambda_1 = 1) = P(\Lambda_1 = 4) = \frac{1}{2}$$

$$X_3 \sim \exp(\Lambda_2), \quad P(\Lambda_2 = 1) = P(\Lambda_2 = 5) = \frac{1}{2}.$$

Obviously $X_1 \supset X_2 \supset X_3$. However the expected makespan of the schedule X_1, X_2, X_3 is 8421/5400 while that of X_1, X_3, X_2 is 8420/5400.

(iii) We were unable to show that SEPT minimizes expected flow time in the exponential case because Lemma 1 only holds for pairwise switches involving the longest remaining task.

REFERENCES

[1] Bruno, J., and P. Downey, "Sequencing Tasks with Exponential Service Times on Two Machines," Technical Report, Department of Electrical Engineering and Computer Science, University of California, at Santa Barbara (1977).

[2] Conway, R.W., W.L. Maxwell and L.W. Miller, *Theory of Scheduling*, (Addison-Wesley, Reading, Mass., 1967).

[3] Karp, R.M., "Reducibility Among Combinatorial Problems," in *Complexity of Computer Computations*, R.E. Miller and J.W. Thatcher (editors), (Plenum Press, 1972).

ON n/1/$\bar{F}$ DYNAMIC DETERMINISTIC PROBLEMS

Ramesh Chandra

University of New Brunswick
Fredericton, New Brunswick, Canada

ABSTRACT

We consider sequencing of *n* jobs which will arrive intermittently and are to be processed on a single machine; the arrival and the processing times of each jobs are assumed known. A schedule is to be developed that minimizes the mean flow time. Two models are considered: (i) when no pre-emption or inserted idle time is allowed in the schedule, and (ii) when pre-emption is allowed but the jobs follow a pre-empt-repeat discipline. We illustrate that Cobham's and Phipp's SPT dispatching rule does not guarantee the optimum $\bar{F}$ even for the non-preemptive model. We propose a branch and bound algorithm for both models and discuss our computational experience. We also examine the relative performances of the optimum nonpre-emptive sequence, and the optimum pre-empt-repeat sequence over that resulting from SPT dispatching rule on a large number of sets of jobs of varying sizes and tightness.

1. INTRODUCTION

We consider the following two models of the basic single machine dynamic deterministic problem:

Model A: a nonpre-emptive n/1/$\bar{F}$,
Model B: pre-empt-repeat n/1/$\bar{F}$.

Model A may be characterized by the following six conditions:

1. There are *n* jobs to be processed. Job j ($j = 1, 2, \ldots, n$) arrives at time r_j and requires p_j units of processing-time. The jobs are numbered as they arrive such that $r_i \leqslant r_j$ if $i < j$. The total number of jobs (n), all r_j's, and all p_j's are fixed and known at the time of scheduling.

2. There is only one machine available. All the *n* jobs must be processed on this machine. The machine cannot handle more than one job at a time but remains available continuously until every job is completed.

3. The processing time of each job is sequence independent. There is no set-up time (or set-up time is included in the processing time), no due date, and no priority attached to any job.

4. No pre-emption and/or inserted idle time is allowed in the schedule. That is, the processing of a job once started cannot be stopped before its completion, and also the machine cannot be kept idle when there is a job waiting to be processed.

5. Define C_j to be the completion time of job j; $c_j \geqslant p_j + r_j$. Also define F_j to be the follow time of job j; $F_j = C_j - r_j$.

6. The solution of the problem involves obtaining a schedule that meets the conditions 1 to 4 and minimizes the mean flow time ($\bar{F}$):

$$\bar{F} = \sum_{j=1}^{n} F_j/n = \sum_{j=1}^{n} (c_j - r_j)/n.$$

In model B condition #4 of model A is relaxed. In this model a job may be pre-empted any number of times but it is assumed that the benefit of any processing that has been done on the job is completely forfeited with every interruption, so that the processing on a pre-empted job must start from the beginning whenever it returns to the machine.

The static $n/1/\bar{F}$ problem, in which all the jobs arrive simultaneously, has a straight forward solution. It is known that there is no need to consider any pre-emption or inserted idle time in the schedule [4]. The mean flow time is minimized by the Shortest Processing Time (SPT) rule which sequences the jobs such that:

$$p_{[1]} \leqq p_{[2]} \leqq \cdots \leqq p_{[n]}$$

where $[j]$ denotes the job occupying the j^{th} position in the sequence.

The dynamic single machine model has been extensively studied in the stochastic form. For the dynamic stochastic model without pre-emptive and inserted idle time features, the SPT dispatching rule is optimal [3,5], (also see [4, p. 166]). This rule selects a new job only after the processing on the job occupying the machine is finished. The next job selected is always from the queue requiring the smallest processing time.

In the above model if pre-emption is permitted and the jobs may be processed in pre-empt resume mode, the "Shortest Remaining Processing Time (SRPT)" rule is optimal [7]. According to this rule, when a job is to be selected from among those waiting, the one with the lowest remaining processing time is chosen. In addition, an arriving job pre-empts the job being processed if the processing time of the new arrival is smaller than the remaining processing time of the job occupying the machine. The problem of minimizing the mean flow time in dynamic stochastic model with pre-empt-repeat feature has not been solved satisfactorily [4].

Consider now extending the results of stochastic version to the deterministic version of the dynamic models when n, r_j, and $p_j (j = 1, 2, \ldots, n)$ are all fixed and known at the time of scheduling. When pre-emption is permitted and jobs may be processed in pre-empt-resume mode, the SRPT dispatching rule sequence is still optimal [6]. But it is easy to see that for nonpre-emptive and noninserted time mode of processing the SPT dispatching rule does not always lead to an optimal sequence. For example consider the following four-job problem:

Job No. (j)	1	2	3	4
Ready time (r_j)	0	0	15	16
Processing time (p_j)	12	15	6	1

The SPT dispatching rule produces the sequence 1-2-4-3 and the mean flow time for this sequence is 17·5. On the other hand, the mean flow time is equal to 15·25 for the sequence 2-3-4-1 and this is the optimal nonpre-emptive sequence. The optimal pre-empt-repeat sequence is 2-4-3-1 which gives the mean flow time equal to 14·75. Thus it may be concluded that the SPT dispatching may be worse than the optimal nonpre-emptive schedule without inserted idle time which in turn may be worse than the optimal pre-empt-repeat schedule, i.e. the schedule with inserted idle time feature. (In a deterministic pre-empt-repeat n/1/$\bar{F}$ problem there is no need to consider any pre-emption. The consequence of a pre-emption may be incorporated by the provision of an inserted idle time in the schedule [4, p. 96], and therefore, the permutation schedules constitute a dominant set. The choice of a permutation schedule uniquely determines the allocation of the inserted idle times [1, p. 83].) However, it is not known what sequencing rule will provide an optimum sequence for either nonpre-emptive n/1/$\bar{F}$ model (model A) or pre-empt-repeat n/1/$\bar{F}$ model (model B).

It is easy to see that one could evaluate every permissible permutation of n jobs to find the best solution for either model A or model B. The real problem is to find either a direct solution or an algorithm that reduces the required number of searches to a computationally practical level. In this paper we consider a general purpose technique, namely, a branch and bound solution algorithm that satisfies the later requirement. Since with only two appropriate changes our algorithm for model A can be adapted to model B, we shall present a common description of the algorithm for both models. (The differences in the algorithm for models A and B are noted in Properties 1 and 2 given later. Properties 1A and 2A are appropriate when dealing with model A while those for model B are 1B and 2B.)

In the implementation of a general purpose technique it is desirable to take advantage of the structure of the problem, which may reduce some computational work. We note below one such property for each of our two models:

PROPERTY 1A: In an n/1/$\bar{F}$ deterministic dynamic problem without pre-emption and inserted idle time features, the first m jobs may be sequenced independently of the remaining n-m jobs where m is the first j for which

$$r_1 + \sum_{i=1}^{j} p_i \leqq r_{j+1};\ j = 1, 2, \ldots, n-1$$

PROPERTY 1B. In an n/1/$\bar{F}$ deterministic dynamic problem with pre-empt-repeat feature, the first m jobs may be sequenced independently of the remaining n-m jobs where m is the first j for which

$$r_j + \sum_{i=1}^{j} p_i \leqq r_{j+1};\ j = 1, 2, \ldots, n-1$$

Property 1A follows directly from the fact that the first m jobs must be processed before the m + 1st job could start since no pre-emption and/or inserted idle time is allowed in the schedule. The legitimacy of Property 1B is argued on the ground that even if the processing of the first m jobs were to start at r_m (and the machine is kept idle during $r_m - r_i$), all the available jobs at r_m may be processed according to the the static SPT rule before the m + 1st job becomes available for processing. Whenever possible, the Property 1A or 1B (for model A or B respectively) may be utilized to partition the original set of jobs into two or more independent subsets of smaller sizes and thereby reduce the computational work.

In section 2 we develop our branch and bound algorithm. In section 3 we present our computational experience in two parts; we first examine the efficiency of our algorithm for both models, and then compare the performance of the SPT dispatching rule schedules with the optimum nonpre-emptive n/1/$\bar{F}$ schedules and the optimum pre-empt-repeat n/1/$\bar{F}$ schedules. Our observations are summarized in Section 4.

2. A BRANCH AND BOUND SOLUTION ALGORITHM

The general branch and bound formulation for a single machine problem is described by Baker in [1]. The solution algorithm that we present below is similar to that of Baker and Su in [2] where they have considered minimizing the maximum tardiness in a pre-empt-repeat deterministic dynamic n/1 model. Each node at level k corresponds to a partial permutation containing k jobs. Associated with each node is a lower bound on the value of the minimum mean flow. This bound is computed by taking the schedule corresponding to the partial sequence and completing the remaining jobs according to SRPT dispatching rule. It follows from [6] that no feasible solution of the given partial sequence can attain a value of $\bar{F}$ less than this bound.

The calculation of lower bounds allows the algorithm to enumerate many permutations only implicitly. If the lower bound of some partial sequence is greater than or equal to the mean flow of a known feasible sequence, it is not necessary to complete the partial sequence in the search for an optimum solution and, therefore, this node must be eliminated from the active list. Define a node to be active if the associated lower bound with this node is less than the mean flow time of the best known feasible sequence. Thus only active nodes are to receive further consideration in the search for an optimum sequence. Since a sequence resulting from SPT dispatching rule is feasible and has the mean flow time, generally very close to the optimum value, the algorithm starts with the SPT sequence as the initial known feasible sequence and its mean flow time, $\bar{F}_{SPT}$, as the known solution.

The branch and bound algorithm maintains a list of all active nodes. At each stage, the node corresponding to the partial sequence with the minimum lower bound is removed and replaced on the list by several nodes corresponding to augmented partial sequences. These are formed by appending one more unscheduled job to the removed partial sequence. The algorithm terminates when there is no active node left on the list.

In order to reduce the computational requirements, in addition to the bounding technique, the algorithm uses two other mechanisms. First, if the SRPT dispatching rule sequences the unscheduled jobs of a node without involving any pre-emption, a feasible sequence is obtained. If the node under consideration is the first one, clearly an optimum sequence has been reached [6], and the algorithm terminates. Otherwise, this solution either displaces the currently known best solution or is inferior to the latter. In any case, there is no further branching from this node.

The second mechanism follows from the requirements and structure of the models. When a partial sequence containing k jobs is removed from the list, it would normally be replaced by $(n\text{-}k)$ augmented sequences, but some of these may be eliminated by taking advantage of Property 2A and 2B for models A anb B respectively.

PROPERTY 2A. In an n/1/$\bar{F}$ deterministic dynamic problem without any pre-emptive and inserted idle time features, if $C_{[k]} < r_n$ and there are $s(s<n-k)$ unscheduled jobs available for processing at $C_{[k]}$ then the removed partial sequence from the list can be replaced by no more than s augmented sequences.

PROPERTY 2B. In an n/1/$\bar{F}$ deterministic dynamic problem with pre-emptive repeat features if

(i) $C_{[k]} < r_n$,

(ii) among the available unscheduled jobs at $C_{[k]}$, job i has the smallest processing time p_i, and

(iii) there are s unscheduled jobs available before $C_{[k]} + p_i$

then the removed partial sequence from the list can be replaced by no more than s augmented sequences.

Property 2A (for model A) follows from the requirement in model A that there be no inserted idle time in the schedule. If there are only s jobs available for processing at $C_{[k]}$ clearly one cannot have more than s augmented sequences without incorporating some inserted idle time in the schedule. Property 2B (for model B) follows from the fact that if the removed partial sequence were to be augmented with a job arriving at or after $C_{[k]} + p_i$, processing of job i could certainly be completed in the k+1st position of the sequence before the arrival of the former.

The s calculated in Property 2A and 2B is in fact the upper bound upon the actual number of augmented sequences resulting from the removed sequence. The actual number will generally be less than s because of the first mechanism discussed earlier.

Because the bound obtained with SRPT dispatching is not too far from the optimal value, and also beçause the upper bound provided by SPT dispatching rule is close to the optimal value, the substitution and elimination mechanisms are very effective in reducing the computational efforts.

3. COMPUTATIONAL EXPERIENCE AND RESULTS

The branch and bound algorithm was implemented as a FORTRAN program on IBM/370/158 Computer of UNB. The algorithm was tested for 240 different problems for each of the two models A and B. These problems were specially designed to explore the algorithm performance by varying the problem size and the tightness of the schedule.

Three problem sizes were examined: n = 10, 20, and 30. Integer values from two different uniform distributors were sampled to generate a job-set constituting a test problem: the arrival times were sampled from a uniform distribution between 0 and $10 \cdot n$; the processing times were independent of the arrival times and were sampled from another uniform distribution between 1 and $20 \cdot \rho$ where ρ is the traffic density. Eight different values of ρ were included: ρ = 0.6, 0.75, 0.9, 0.1, 1.25, 1.5, 2.0, and 5. For each n and ρ combination 10 different job-sets were randomly generated. Thus, in all 240 different sets of jobs were generated to obtain the 240 test problems.

The same 240 problems were sequenced under both model A and model B. Properties 1A and 2A were incorporated into the algorithm when sequencing the jobs under model A; when sequencing under model B properties 1B and 2B were used. The algorithm obtained an optimum sequence for each of the 240 problems under each of the two models.

The entire computation was completed in six computer runs; three runs for each of the two models, and one run for each problem size. Each run involved sequencing 80 problems resulting from ten replications for each of the eight ρ's. The computational results for both models A and B are summarized in Tables I, II, and III for n = 10, 20, and 30 respectively.

The average CPU time/problem (including the time required for generating the problems) under model A were 0·078, 1·808 and 76·562 seconds for n = 10, 20, and 30 respectively. The corresponding figures under model B were 0.089, 1.427, and 92.675. These figures indicate that our algorithm is highly efficient for both models A and B. However, the trend clearly indicates that as n increases the required CPU time increases exponentially.

We now compare the performance of the SPT dispatching rule sequence with the optimum nonpre-emptive sequence and the optimum pre-empt-repeat sequence. The optimum total flow time for 10 problems for each of 24 n and ρ combination under SPT dispatching rule, under nonpre-emptive model, and under pre-empt-repeat model are given in Tables I through III in Columns 1, 2, and 5 respectively. The relative improvement in flow by the optimum nonpre-emptive schedule over the SPT dispatching rule is shown in Column 4. The relative improvement by the optimal pre-empt-repeat schedule over the SPT dispatching rule is shown in Column 8. The entries in Columns 4 and 8 clearly indicate that the improvement in the mean flow time by either the optimum nonpre-emptive schedule or the optimum pre-empt-repeat schedule over the SPT dispatching rule sequence is very little if any.

TABLE I. *Total Flow and Computational Experience*
(Cumulative results for 10 problems for each ρ; $n = 10$)

S.N.	ρ	SPT SEQ.	Nonpre-emptive Optimal SEQ.			Pre-emptive-repeat Optimal Sequence			
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Total Flow	Total Flow	No. of nodes generated	% improvement $\frac{(1)-(2)}{(2)}$	Total flow	No. of nodes generated	Total inserted idle time	% improvement $\frac{(1)-(5)}{(5)}$
1	0.6	1035	1035	63	0	1028	90	5	0.680
2	0.75	1287	1287	92	0	1280	176	3	0.546
3	0.9	1896	1896	119	0	1882	258	16	0.850
4	1.0	1766	1759	175	0.397	1738	294	18	1.611
5	1.25	3123	3114	175	0.389	3071	294	23	1.693
6	1.5	3612	3612	109	0	3590	199	4	0.612
7	2.0	5101	5090	146	0.216	4968	307	33	2.677
8	5.0	19395	19395	51	0	18477	176	34	4.968
Total		37217	37188	930	0.077	36034	1794	136	3.289

	Nonpre-emptive	Pre-emptive repeat
Total CPU time for 80 problems (in seconds on IBM 370/158)	6.26	7.14
Time/problem	0.078	0.089
No. of nodes/problem	12	22

Further, in most practical stituations all p_j's and r_j's have to be estimated. These estimates themselves are quite liable to be imprecise. In light of the fact that the differences between $\bar{F}_{\text{pre-empt-repeat}}$, $\bar{F}_{\text{Nonpre-emptive}}$, and $\bar{F}_{\text{SPT}}$ are very small even in the deterministic case, an optimal schedule (with or without inserted idle time feature) based on imprecise estimates can hardly be expected to preform significantly better than the SPT dispatching rule sequence based on the same estimates.

4. CONCLUSIONS

We have developed a branch and bound algorithm for minimizing the mean flow time of a "dynamic deterministic n jobs one machine problem" when: (i) pre-emption and inserted idle

TABLE II. *Total Flow and Computational Experience*
(Cumulative results for 10 problems for each ρ; n = 20)

S.N.	ρ	SPT SEQ.	Nonpre-emptive Optimal SEQ.			Pre-emptive-repeat Optimal Sequence			
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Total Flow	Total Flow	No. of nodes generated	% improvement $\frac{(1)-(2)}{(2)}$	Total flow	No. of nodes generated	Total inserted idle time	% improvement $\frac{(1)-(5)}{(5)}$
1	0.6	82153	2153	426	0	2132	741	8	0.984
2	0.75	2915	2901	782	0.480	2863	1472	14	1.810
3	0.9	4563	4559	2925	0.087	4436	3805	27	0.608
4	1.0	4667	4635	8373	0.258	4489	6538	41	3.519
5	1.25	7310	7293	3620	0.233	7232	4970	22	1.078
6	1.5	11165	11137	3810	0.251	11023	3916	18	1.288
7	2.0	18380	18316	6190	0.349	18244	7534	14	.745
8	5.0	63663	63638	1055	0.039	62693	1111	13	1.547
Total		114796	114632	27181	0.039	113112	30087	157	1.488

	Nonpre-emptive	Pre-emptive repeat
Total CPU time for 80 problems (in Seconds on IBM 370/158)	144.66	114.18
Time/problem	1.808	1.427
No. of nodes/problem	340	376

TABLE III. *Total Flow and Computational Experience*
(Cumulative results for 10 problems for each ρ; n = 30)

S.N.	ρ	SPT SEQ.	Nonpre-emptive Optimal SEQ.			Pre-emptive-repeat Optimal Sequence			
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Total Flow	Total Flow	No. of nodes generated	% improvement $\frac{(1)-(2)}{(2)}$	Total flow	No. of nodes generated	Total inserted	% improvement $\frac{(1)-(5)}{(5)}$
1	0.6	3082	3079	236	0.096	3041	1160	16	1.348
2	0.75	4426	4417	3585	0.203	4407	12338	3	0.431
3	0.9	6389	6352	7216	0.582	6249	12573	30	2.240
4	1.0	9670	9628	48369	0.436	9609	64635	12	0.634
5	1.25	19763	18680	178762	0.337	18667	142348	10	0.417
6	1.5	25570	25546	94138	0.094	25458	52859	13	0.439
7	2.0	39894	39807	12643	0.218	39651	19864	3	0.612
8	5.0	117177	117143	13691	0.029	115730	9315	67	1.250
Total		224951	224652	358640	0.133	222812	314102	154	0.960

	Nonpre-emptive Sequencing	Pre-emptive repeat Sequencing
Total CPU time for 80 problems. (in seconds on IBM 370/158	6125.02	7414.35
Time/problem	76.562	92.675
No. of nodes/problems	2785	3938

time are not allowed, or (ii) pre-emption is allowed but the jobs follow a pre-empt-repeat discipline. Our algorithm is quite efficient for moderate n ($n \leqslant 30$).

Our results show that the optimum nonpre-emptve or pre-empt-repeat $n/1/\bar{F}$ schedules are not significantly superior to Phipp's SPT dispatching rule schedules.

REFERENCES

[1] Baker, K. R. *Introduction to Sequencing and Scheduling*, (John Wiley & Sons Inc., 1974).

[2] Baker, K. R. and Z. Su, "Sequencing with Due-dates and Early Start Times to Minimize Maximum Tardiness," Naval Research Logistics Quarterly 20, No. 1 (March 1974).

[3] Cobham, A. "Priority Assignment in Waiting Line Problems," Management Science 10, No. 1 (October 1963).

[4] Conway, R. W., W. Maxwell, and R. Miller, *Theory of Scheduling*, (Addison-Wesley, Reading, Mass., 1967).

[5] Phipps, T. E. Jr., "Machine Repair as a Priority-Waiting-Line Problem," Operations Research 4, No. 1 (February 1956).

[6] Schrage, L. E. "A Proof of the Optimality of the Shortest Remaining Processing Time Discipline," Operations Research 6, No. 3 (May-June 1968).

[7] Schrage, L. E. and L. W. Miller, "The Queue M/G/1 with the Shortest Remaining Time Discipline," Operations Research 14, No. 3 (1966).

[8] Smith, W. E. "Various Optimizers for Single-State Production," Naval Research Logistics Quarterly 3, No. 1 (March 1956).

DAMAGE CALCULATIONS FOR UNRELIABLE WARHEADS

R. T. Curran, S. C. Jaquette, and J. L. Politzer

Systems Control, Inc.
Palo Alto, California

ABSTRACT

For large numbers of perfectly reliable, optimally targeted warheads the square-root law approximates the expected fraction damage achieved on an area target. In this paper a more exact expression is derived for this damage fraction which holds for all numbers of warheads. This expression is shown to converge to the square-root law when a large number of warheads are fired. The more exact expression is used in a procedure to calculate expected damage when warheads are unreliable, and this procedure is shown to be superior to a modified square-root approximation which has been used previously.

INTRODUCTION

A key problem in analyzing the effects of a salvo attack against an area target is to estimate the damage that will result. This problem is important in a military engagement such as a missile attack on a city or an anti-aircraft attack on an air-squadron for both the offense and defense. The offense will seek a strategy for attacking the area target to maximize damage; the defense will seek strategies to minimize damage. Both purposes will be served by developing an exact expression for the expected damage that will occur when a salvo of optimally-targeted but unreliable warheads is sent against an area target whose relative value within the entire target area is a known function of position within the target area. Approximations to the expected damage expression will also be developed and examined to indicate the limits of applicability.

An historical perspective of this problem can be found in Ref. [5]. The starting point for our purposes is the asymptotic expected damage expression for a perfectly reliable salvo attack on an area target. For a large number of reliable bursts optimally targeted to an area target whose value function has a symmetric bivariate (or circular) normal distribution, the damage has been calculated using the square-root damage law:

$$E_N = 1 - (1 + K\sqrt{N})\, e^{-K\sqrt{N}} .$$

N is the number of warheads sent and K is a factor associated with the target which combines the effects of target hardness and size and warhead and delivery system characteristics. The principal problem addressed in this note is to find an exact expression for E_N for an arbitrary (small) number of unreliable warheads when K is defined in agreement with the square-root law and the target has symmetric normally distributed value. This is an alternative to inaccurate approximations such as the K' method presented in Ref. [4] using the square-root law. A derivation of the square-root law is also given, as existing derivations are inaccessible or quite abstract (see Refs. [1] and [3]). The development below is based on work in Ref. [5] and corrects the extension of Ref [5] found in Ref. [2].

2. DEVELOPMENT OF AN EXACT EXPRESSION OF E_N

Walsh [5] derives the probability density function, P, of weapon impact points to maximize target damage for an arbitrary target, where the relative value of points within the target area is given by a probability density function, i.e., the total target value is normalized to 1. To the degree that warheads can be targeted and delivered according to the probability distribution approximated by P, Walsh's expression for E_N represents the damage from optimally targeted (maximum damage) reliable warheads. In every case his E_N is an upper bound to attainable damage. Walsh's results expressed in polar coordinates are:

$$(1) \qquad P(r,\theta) = \begin{cases} 1 - \left(\dfrac{\lambda}{T(r,\theta)}\right)^{\frac{1}{N-1}} & ; T(r,\theta) > \lambda \\ 0 & ; T(r,\theta) \leqslant \lambda \end{cases}$$

$$(2) \qquad E_N = \iint_{T(r,\theta)>\lambda} \left\{ T(r,\theta) - \left(\frac{\lambda^N}{T(r,\theta)}\right)^{\frac{1}{N-1}} \right\} r dr d\theta$$

where N is the number of warheads delivered to a target centered at $(r,\theta) = (0,0)$ with relative value of the point (r,θ) given by the density function $T(r,\theta)$ and where λ (a Lagrange multiplier entering the derivation) is the solution of

$$(3) \qquad \iint_{T(r,\theta)>\lambda} \left\{ 1 - \left(\frac{\lambda}{T(r,\theta)}\right)^{\frac{1}{N-1}} \right\} r dr d\theta = \pi R^2 .$$

The right hand side of (3), πR^2, is the area destroyed by a single burst.

For the symmetric Gaussian valued target assumed for most of the subsequent development,

$$T(r,\theta) = \frac{1}{2\pi\sigma^2} e^{-\frac{r^2}{2\sigma^2}} .$$

where σ is a scale factor for the absolute size of the target. Hence the region of integration in (3), where T will be greater that λ, is a circle of radius r_o, where

$$r_o = \sqrt{-2\sigma^2 ln \lambda 2\pi\sigma^2} .$$

Note that this expression, and (3), only have meaning when $\lambda < (2\pi\sigma^2)^{-1}$ Equations (1) through (3) can be simplified by performing the integration on θ, where $T(r) \equiv T(r,\theta)$ is independent of θ. Using the expression for $T(r)$ and integrating Equations (2) and (3), the following are obtained:

(4)
$$E_N = \begin{cases} 1 - \exp\left\{-\dfrac{R^2}{2\sigma^2}\right\} & ; N = 1 \\ 1 - 2\pi\lambda\sigma^2 N\left[1 - \dfrac{N-1}{N}\,(2\pi\lambda\sigma^2)^{\frac{1}{N-1}}\right] & ; N \neq 1 \end{cases}$$

(5)
$$\lambda = \begin{cases} \text{for } N = 1\text{: } \dfrac{1}{2\pi\sigma^2}\exp\left\{\dfrac{-R}{2\sigma^2}\right\} \\ \text{the solution to the following equation for } N \neq 1\text{:} \\ (2\pi\lambda\sigma^2)^{\frac{1}{N-1}}\exp\left\{-(2\pi\lambda\sigma^2)^{\frac{1}{N-1}}\right\} = \exp\left\{-\left[1 + \dfrac{R^2}{2\sigma^2(N-1)}\right]\right\} \end{cases}$$

Equation (5) may be solved for λ fairly easily using Newton-Raphson iteration. Let $Q = (2\pi\lambda\sigma^2)^{\frac{1}{N-1}}$, then (5) becomes

(6)
$$Qe^{-Q} = \exp\left\{-\left[1 + \frac{R^2}{2\sigma^2(N-1)}\right]\right\}.$$

Equation (6) may be solved iteratively by a recursion which converges quickly for all R, σ, and N: take

$$Q_o = \exp\left\{-\left[1 + \frac{R^2}{2\sigma^2(N+1)}\right]\right\} \text{ and iterative recursion } Q_{i+1} = \left[\frac{Q_i}{Q_i - 1}\right]\left[1nQ_i + \frac{R^2}{2\sigma^2(N-1)}\right].$$

With R/σ identified as K in the square-root damage law, an algorithm to calculate E_N exactly may be summarized as follows:

For $N = 1$; $E_1 = 1 - e^{-\frac{K^2}{2}}$

for $N \neq 1$; $E_N = 1 - NQ^{N-1} + (N-1)Q^N$, where Q is found by the iterative recursion

$$Q_0 = \exp\left\{-1 - \frac{K^2}{2(N-1)}\right\}$$

and

$$Q_{i+1} = \left[\frac{Q_i}{Q_i - 1}\right]\left[\frac{K^2}{2(N-1)} + lnQ_i\right],\ (i > 0)\ .$$

The above algorithm will calculate Q to within 0.0001 in at most 14 iterations. Convergence can be speeded considerably if an alternative Q_o is used for small values of $K^2/(N-1)$, i.e., $Q_0 = 1 - \dfrac{K}{\sqrt{N}}$.

3. DERIVATION OF SQUARE-ROOT DAMAGE LAW FOR RELIABLE WARHEADS

The square-root damage law can be derived from the exact theory as an asymtotic approximation. As such it can be expected to approximate the exact expression for large values of N.

Taking logarithms of (5) and setting $R/\sigma = K$, we obtain

$$Q - lnQ = 1 + \frac{K^2}{2(N-1)} .$$

For $N > > 1$, Q is near 1, and the logarithm can be approximated by the first two terms of its Taylor expansion $lnQ \approx (Q-1) - \frac{(Q-1)^2}{2} + \ldots$ In this form the following approximations are obtained:

$$Q \approx 1 - \frac{K}{\sqrt{N-1}} \quad \text{and} \quad E_N \approx 1 - Q^{N-1}(1 + K\sqrt{N-1}) .$$

In taking the limit as N gets large, some normalization is needed, otherwise E_N will converge to 1. Assume that the total destructive power of the attack remains constant but that it can be generated in any number of equally sized weapons, i.e., this means that $(N\pi R^2)/(\pi\sigma^2)$ remains constant as N increases or equivalently that $K\sqrt{N} = (R/\sigma)\sqrt{N}$ will remain constant as N increases. With this normalizing assumption, recalling that

$$e^x = \lim_{n\to\infty} (1 + \frac{x}{n})^n .$$

and rewriting Q^{N-1} as

$$Q^{N-1} = \left|1 - \frac{k}{\sqrt{N-1}}\right|^{N-1} = \left|1 - \frac{K\sqrt{N-1}}{N-1}\right|^{N-1} ,$$

one obtains the result that $Q^{N-1} \approx \exp\{-K\sqrt{N}\}$ for large N. Hence,

$$E_N \approx 1 - (1 + K\sqrt{N})e^{-K\sqrt{N}} \quad \text{for large } N .$$

The exact damage law can easily be compared with the square-root law for different K and N. The square-root law always underestimates the damage. The two calculations tend to agree asymptotically for large N, and they are no more than 3% apart for $N > 7$, and $0.6 < K < 4$.

The square-root law could be modified slightly to produce much better empirical agreement with the more exact calculaton. This modified square-root law is:

(7)
$$E_1 = 1 - e^{-\frac{K^2}{2}}$$
$$E_N = 1 - (1 + K\sqrt{N+1})e^{-K\sqrt{N+1}} \quad (N>1) .$$

The difference between this and the exact law is less than 4%, and less than 3% for $N > 3$. In fact for $K = 1.14$, the modified square-root law at $N = 2(3)$ is 2% (1%) too low and less than 1% too low for all other values of N.

4. EXPECTED DAMAGE WHEN WARHEADS ARE UNRELIABLE

The expressions for E_N as given by any of the previous formulas give the expected damage given that exactly N optimally targeted warheads are sent and detonated. In the general case there is a probability, P_N, that exactly N warheads out of the total of M sent do indeed succeed, where the P_N may be calculated from a simulation of an entire engagement or other method. The total expected damage, E, is given properly by

$$E = \sum_{1}^{M} E_N P_N$$

under the assumption that the N successful warheads are laid down approximately optimally given that there are N successes. Note that this is not the same as the damage expected from $\sum NP_N$ reliable warheads unless E_N is linear in N. As an approximation, one may assume that each of the warheads has a probability of success P_L. In this case, P_N is binomial, and

$$E = \sum_{N=1}^{M} E_N \binom{M}{N} (P_L)^N (1-P_L)^{M-N} .$$

The last equation may be compared with the method of [4] which uses a modified K value denoted K' in the square root-law to obtain E. In [4] K' is the solution of

$$D \equiv P_L \cdot [1 - (1 + K)e^{-K}] = 1 - (1 + K')e^{-K'} .$$

which may be obtained from the iteration formula:

$$K'_{i+1} = K'_i + [(1 + K'_i) + e^{K'_i}(D-1)]K'_i, \text{ with } K_0 = 1 .$$

This method [4] then estimates the total expected damage by

$$E = 1 - (1 + K'\sqrt{N})e^{-K'\sqrt{N}} . \tag{8}$$

The method uses an approximation to the proper expected value and an approximation to the exact damage. Thus (8) cannot be expected to be accurate for all values of K and N.

The four different approaches discussed above are illustrated in Figure 1 for three values of K. These show that E computed correctly using the exact theory for E_N is greater than E computed using the square-root law and taking the expected value properly. Both of these are greater than E using the K' approach. The empirical correction to the square-root law given in (7) is also displayed assuming the expected value for unreliable warheads is taken correctly.

It is clear from these results that the exact theory for E_N using (4) and (5) and the correct expected value expression for E is preferred, although the $N+1$ square root expression (7) and a proper expectation is a possibly acceptable alternative. Both of these are superior to the K' method (8) and the usual square-root law with proper expected value calculation.

REFERENCES

[1] Duncan, R. L., "Hit Probabilities for Multiple Weapons Systems," SIAM Review, 6, 111-114 (1964).

[2] Eckler, A. R. and S. A. Burr, *Mathematical Models of Target Coverage and Missile Allocation*, Military Operations Research Society (1972).

[3] Galiano, R. J. and H. Everett, "Defense Modeis IV," Paper 6. Lambda Corp., Arlington, VA (1967).

[4] Kopp, R. G. and J. T. Steinberg, "K Factor Adjustments for Weapon Yield and Reliability," Lockheed Missiles and Space Co., Inc., Sunnyvale, CA (1975).

[5] Walsh, J. E., "Optimum Ammunition Properities for Salvos," Operations Research, *4*, 204-212 (1956).

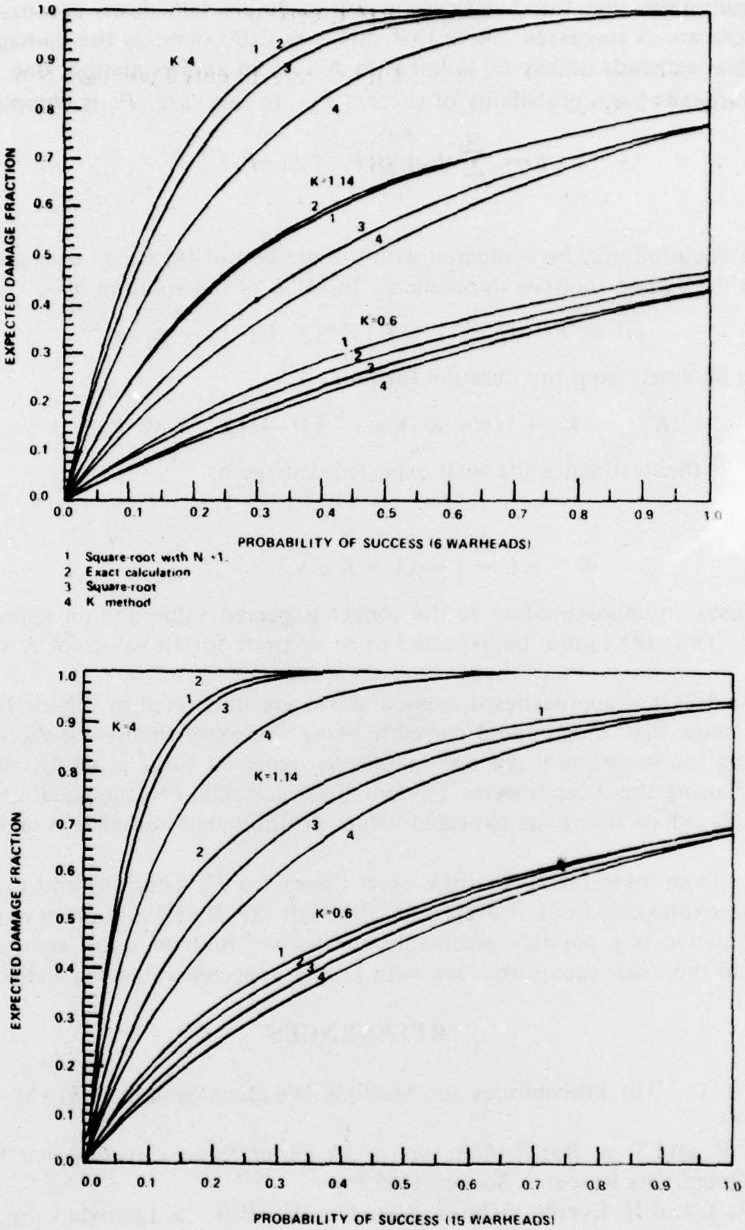


FIGURE 1. Comparison of damage formulae

CORRIGENDUM

The following listing was omitted from the Cumulative Twenty Five Year Index which appeared in the December 1978 issue:

> Love, R. F., "A Two-Station Stochastic Inventory Model with Exact Method of Computing Optimal Policies," Vol. 14, No. 2, June 1967, pp. 185-217

☆ U. S. GOVERNMENT PRINTING OFFICE: 1979 — 281-491/4